Helsinki University Biomedical Dissertations No. 193

Computational methods for analyzing complex high-throughput data from cancers

Kristian Ovaska

Research Programs Unit,
Genome-Scale Biology Research Program
Institute of Biomedicine,
Biochemistry and Developmental Biology
Faculty of Medicine
University of Helsinki
Finland

Academic dissertation

To be publicly discussed, with the permission of
the Faculty of Medicine of the University of Helsinki,
in Biomedicum Helsinki 1, Lecture Hall 2, Haartmaninkatu 8, Helsinki
on April 25th, 2014, at 12 o'clock noon.

Helsinki 2014

Supervisor

Sampsa Hautaniemi, DTech, Professor
Genome-Scale Biology Research Program
Institute of Biomedicine
Faculty of Medicine, University of Helsinki
Helsinki, Finland

Reviewers

Matti Nykter, PhD, Professor
Institute of Biomedical Technology, University of Tampere
Tampere, Finland

Jorma Palvimo, PhD, Professor
Institute of Biomedicine, University of Eastern Finland
Kuopio, Finland

Official opponent

Robert Clarke, PhD, DSc, Professor
Department of Oncology, Georgetown University
Washington, D.C., USA

Helsinki University Biomedical Dissertations No. 193
ISSN 1457-8433

ISBN 978-952-10-9816-1 (paperback)
ISBN 978-952-10-9817-8 (PDF)
http://ethesis.helsinki.fi
Unigrafia Oy
Helsinki 2014

Contents

Publications and author's contributions

Publication I **Ovaska K**, Laakso M, Haapa-Paananen S, Louhimo R, Chen P, Aittomäki V, Valo E, Núñez-Fontarnau J, Rantanen V, Karinen S, Nousiainen K, Lahesmaa-Korpinen A-M, Miettinen M, Saarinen L, Kohonen P, Wu J, Westermarck J, Hautaniemi S. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Medicine*, 2010, 2:65.

Publication II **Ovaska K**, Matarese F, Grote K, Charapitsa I, Cervera A, Liu C, Reid G, Seifert M, Stunnenberg HG, Hautaniemi S. Integrative analysis of deep sequencing data identifies estrogen receptor early response genes and links *ATAD3B* to poor survival in breast cancer. *PLOS Computational Biology*, 2013, 9(6): e1003100.

Publication III **Ovaska K**, Lyly L, Sahu B, Jänne O, Hautaniemi S. Genomic region operation kit for flexible processing of deep sequencing data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2013, 10(1): 200-206.

Author's contributions

Publication I Main design and implementation of the Anduril framework; implementation of 80 analysis components for Anduril (some together with other authors); analyzing expression, DNA methylation and survival data from glioblastoma; integrating TCGA analysis results into a WWW site; drafting the manuscript.

Publication II Design and implementation of the SPINLONG method; analysis of PolII, histone alteration and ER ChIP sequencing data from MCF-7; survival analysis; drafting the manuscript.

Publication III Design of the region algebra; design and implementation of GROK (together with LL); analyzing ChIP sequencing data from LNCaP cells; drafting the manuscript.

Related publication

Policies of the Medical Faculty at the University of Helsinki limit the use of publications in multiple doctoral theses of different individuals. The following publication is thus included as a "related publication".

Related Publication I Sahu B, Laakso M, **Ovaska K**, Mirtti T, Lundin J, Rannikko A, Sankila A, Turunen JP, Lundin M, Konsti J, Vesterinen T, Nordling S, Kallioniemi O, Hautaniemi S, Jänne O. Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signaling and prostate cancer. *EMBO Journal*, 2011, 30(19): 3962-3976.

Author's contributions to related publication

Related Publication I Analysis of histone modification ChIP sequencing and DNaseI hypersensitivity data; contibutions to figures in manuscript.

Abbreviations

AI	Artificial intelligence
ALT	Alternative pathway for telomere elongation
AR	Androgen receptor
ARBS	Androgen receptor binding site
ATAD3B	ATPase Family, AAA Domain Containing 3B
BAM	(Binary) Sequence Alignment/Map format (file format)
BLAST	Basic Local Alignment Search Tool (method)
BRCA(1/2)	Breast Cancer 1/2, early onset
BS	Binding site
BWA	Burrows-Wheeler Alignment (method)
CDH1	Cadherin 1, type 1, E-cadherin
CDKN2A	Cyclin-dependent kinase inhibitor 2A
cDNA	Complementary DNA
CGH	Comparative genomic hybridization
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag Sequencing
ChIP-chip	Chromatin immunoprecipitation followed by DNA microarray analysis
ChIP-qPCR	Chromatin immunoprecipitation followed by quantitative polymerase chain reaction
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
ChIP	Chromatin immunoprecipitation
Cy(3/5)	Cyanine dye 3/5
ddNTP	Dideoxynucleotide
DHT	(5α-)Dihydrotestosterone
DNA	Deoxyribonucleic acid
DNaseI	Deoxyribonuclease I
E. coli	*Escherichia coli*
E2F	E2 transcription factor (gene family)
ELAND	Efficient Large-scale Alignment of Nucleotide Databases (method)
EMT	Epithelial-mesenchymal transition
ENCODE	Encyclopedia of DNA Elements (project)
ER	Estrogen receptor (α or β)
FACS	Fluorescence activated cell sorting
FOXA1	Forkhead box A1 (gene or protein)
G_1	Gap 1 phase of mitosis
Gb	Gigabases
GBM	Glioblastoma multiforme
Gbp	Giga base pairs
GRO-seq	Global Run-On Sequencing
GROK	Genomic Region Operation Kit (method)
H2A.Z	Histone H2A variant Z
H3K4me3	Histone H3 lysine 4 trimethylation
HER2	Human epidermal growth factor receptor 2
IBC	Invasive breast cancer
ICGC	International Cancer Genome Consortium (project)
KC	Kolmogorov complexity
KM	Kaplan-Meier (survival analysis)
MACS	Model-based Analysis for ChIP-Seq
MAPK	Mitogen-activated protein kinase
MCF-7	Michigan Cancer Foundation 7 (cell line)
MET	Mesenchymal-epithelial transition

miRNA	Micro RNA
mRNA	Messenger RNA
MSN	Moesin
P	Phosphorylation
PCR	Polymerase chain reaction
PES	Paired-end sequencing
PolII	RNA polymerase II
PPI	Protein-protein interaction
PR	Progesterone receptor
PTM	Post-translational modification
RB1	Retinoblastoma 1
RNA	Ribonucleic acid
S	Synthesis phase of mitosis
SGS	Second-generation (DNA) sequencing
siRNA	Small interfering RNA
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variation
SPINLONG	Spatial Pattern Identification by Non-Linear OptimizatioN with Global constraints (method)
TCGA	The Cancer Genome Atlas (project)
TERT	Telomerase reverse transcriptase
TF	Transcription factor
TSG	Tumor suppressor gene
TSS	Transcription start site
UTR	Untranslated region
VCF	Variant Call Format (file format)
VEGF	Vascular endothelial growth factor
WHO	World Health Organization
WWW	World Wide Web

Abstract

Cancers are a heterogeneous group of diseases that cause 7.6 million deaths yearly worldwide. At the cellular level, cancer is characterized by increased proliferation and invasion of tissue. These phenotypes are caused by environmental or inherited factors that increase the mutability of the genome, leading to dysregulation of a number of cellular processes. Identifying the genotypic changes and their phenotypic consequences is key to accurate diagnosis and prognosis, as well as improved treatment regimens.

Cancer cells can be investigated at a genome-wide scale using high-throughput measurement techniques such as DNA sequencing and microarrays. These rapidly evolving technologies provide experimental data that have two challenging characteristics: the volume of data is large and data are structurally complex. These data need to be analyzed in an accurate and scalable manner to arrive at biomedically relevant conclusions.

I have developed three computational methods for analyzing high-throughput genomic data, and applied the methods to experimental data from three cancers. The first computational method is an extensible workflow framework, Anduril, for organizing the overall software structure of an analysis in a scalable manner. The second method, SPINLONG, is a flexible algorithm for analyzing chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) data from complex experimental designs, such as time series measurements of multiple markers. The third method, GROK, is used for preprocessing deep sequencing data. Its design is based on a mathematical formalism that provides a succinct language for these operations.

The experimental part studies gene regulation and expression in glioblastoma multiforme, and breast and prostate cancer. The results demonstrate the applicability of the developed methods to cancer research and provide insights into the dysregulation of gene expression in cancer. All three studies use both cell line and clinical material to connect the molecular and disease outcome aspects of cancer. These experiments yield results at two conceptual levels. At the holistic level, lists of significant genes or genomic regions provide a genome-wide view into genomic alterations in cancer. At the specific level, we focus on one or a few central genes, which are experimentally validated, to provide an accessible starting point for understanding the results. Together, the thesis focuses on understanding the complexity of cancer and managing the complexity of genome-wide data.

Tiivistelmä

Syövät ovat heterogeeninen joukko sairauksia, jotka aiheuttavat vuosittain 7,6 miljoonaa kuolemaa maailmanlaajuisesti. Solutasolla syövälle on ominaista lisääntynyt solukasvu sekä leviäminen ympäröivään kudokseen. Nämä solutason ilmiöt johtuvat ympäristö- ja perinnöllisistä tekijöistä, jotka lisäävät genomin mutaatioalttiutta ja häiritsevät solun biokemiallisia prosesseja. Syövän hoidolle sekä diagnoosille on tärkeää tunnistaa geneettiset muutokset syöpäsoluissa sekä niiden vaikutukset fenotyyppiin.

Syövän solumuutoksia voi tutkia hiljattain kehitetyillä genominlaajuisilla mittaustekniikoilla, kuten DNA:n sekvensoinnilla ja mikrosiruilla. Nämä uuden sukupolven tekniikat tuottavat mittaustietoa, jolla on kaksi ominaispiirrettä: sitä on määrällisesti paljon ja se on rakenteeltaan monimutkaista. Tällainen mittaustieto on kyettävä analysoimaan täsmällisesti ja laskennallisesti skaalautuvasti, jotta tutkimuksesta saadaan lääketieteellistä lisäarvoa.

Tässä työssä on kehitetty kolme laskennallista menetelmää genominlaajuisten aineistojen analyysiin, sekä hyödynnetty näitä menetelmiä kokeellisesti kolmen syövän tutkimuksessa. Ensimmäinen laskennallinen menetelmä on ohjelmistokehys Anduril, joka tarjoaa laajennettavan työnkulkuihin perustuvan alustan suurten ja monimutkaisten aineistojen analysointiin. Toinen menetelmä on SPINLONG-algoritmi, jolla analysoidaan proteiinien sitoutumista DNA:han genominlaajuisesti. Kolmas menetelmä, GROK, on ohjelmisto laajojen DNA-sekvensointiaineistojen tehokkaaseen esikäsittelyyn.

Työn kokeellinen osuus käsittelee geenien ilmentymistä ja säätelyä glioblastoomassa sekä rinta- ja eturauhassyövässä. Saadut tulokset osoittavat kehitettyjen laskennallisten menetelmien soveltuvuutta kokeelliseen tutkimukseen ja lisäävät tietämystä näissä syövissä tapahtuvista genomitason muutoksista. Kokeellisissa tutkimuksissa on hyödynnetty sekä soluviljelmiä että potilasnäytteitä kytkemään molekyylitason muutokset kliiniseen tulokseen. Kokeista saatuja tuloksia voi tarkastella kahdella abstraktiotasolla. Holistisella tasolla, johon kuuluu listoja muuntuneista geeneistä sekä kromosomialueista, saadaan kokonaiskuva genominlaajuisista muutoksista syövissä. Spesifisellä tasolla tarkennetaan oleellisimpiin geeneihin, joiden merkitys on kokeellisesti todennettu, mikä tarjoaa luontevan lähtökohdan tuloksien tulkintaan. Kokonaisuutena väitöskirja tutkii syövän monimutkaisuutta ja kehittää menetelmiä monimutkaisten genominlaajuisten aineistojen tulkitsemiseen.

1 Introduction

Complex systems are characterized by a large number of interacting components, emergent behavior, adaptability to change, and memory of past events [1]. Such systems include social structures, human cells, economies, and computer systems. For scientific fields studying these systems, a basic challenge is how to observe, interpret and manage the underlying complexity.

Human cells are important examples of complex systems [2]. Their pathological states, such as those observed in cancer [3, 4], are a major societal issue, leading to loss of life and economical costs [5]. The goal of biomedical research is to better observe (diagnose), predict (prognose) and manipulate (treat) the behavior of diseased cells. As is usual with complex systems, implementing such a goal is challenging, and requires efforts both at the experimental as well as the methodological level.

Empirical sciences, such as biomedicine, follow the process of conducting experiments, interpreting the resulting data, drawing conclusions, and altering the current theories. Biological experiments can be divided into low-throughput biology, which observes a small number of variables (e.g., a few genes) concurrently and generates a small number of total data points [6, 7]; and high-throughput biology, which observes a high number of variables (e.g., a genome) with a large number of data points [8, 9]. In this thesis, the focus is on the latter. High-throughput biology is distinct from sciences such as psychology, which generate data with a complex structure but small volume [10], and particle physics, which generates large-scale data with relatively simple structure [11]. A key question in high-throughput biology is how to efficiently conduct experiments whose results are complex in both the volume and structure dimensions.

The two dimensions of complexity in high-throughput biology are addressed in different manners. Processing large volumes of data is predominantly a computer performance issue, with a human contribution in designing efficient algorithms. Analyzing structurally complex data, in contrast, is predominantly a human challenge, requiring the researcher to browse data manually or to use sophisticated programs to derive results.

The goal of this thesis is to develop automated methods for analyzing biomedical high-throughput data by taking both complexity dimensions into account, and to apply these methods to cancer research. The emphasis in method development is on enabling the researcher to work on structurally complex experiments, while also ensuring adequate computational efficiency. The cancer experiments combine high-throughput molecular data with basic phenotypes of patients to derive clinically relevant results.

2 Complexity in molecular biology and system design

In this section, we elucidate the foundations of complex systems theory in the context of molecular biology in order to obtain a schema that binds the themes in this thesis together. Complexity theory both helps understand the cancer systems under study, and suggests strategies for solving technical challenges in method development. This is an illustration of a duality in complexity: it is present both in evolved biological systems and in rationally engineered technological systems.

Complexity can be analyzed and characterized using two basic strategies. The first, qualitative strategy describes common features of complex systems. The second, quantitative strategy aims to measure complexity using a variety of metrics.

2.1 Qualitative features of complex systems

Complex systems have a large number of interacting components [1], such as proteins in a cell or cells in a body [2, 12]. Complexity arises from the quantity of the components, but also from their interactions and heterogeneity. For instance, there are more than 20,000 protein-coding genes in the human genome [13] and 110,000 binary protein-protein interactions (PPIs) in the proteome [14].

Complex systems retain memory of past events, often established using positive feedback systems [1]. For instance, a white blood cell has differentiated from a hematopoietic stem cell and maintains the differentiation state using biochemical feedback mechanisms [15]. Presence of such a system state indicates that the system cannot be fully understood solely by enumerating its components. Cancer cells undergo genetic mutation, enabling plasticity at the genomic level [3]; thus, their internal state is particularly complex [16].

Complex systems exhibit emergent behavior arising from the interactions of its components and their current state [1]. Such behavior can be extreme, i.e., a small perturbation may lead to large, only partially predictable consequences. In cancer biology, a prime example is the presence of germline single nucleotide variations in the *p53* gene in the Li-Fraumeni syndrome, which leads to a significantly increased lifetime cancer risk [17, 18]. One functioning copy of *p53* in a given cell is generally sufficient to suppress cancer, but randomly this functional copy is lost and, due to disrupted interactions between *p53* and other genes, a tumor may develop. Understanding and predicting emergent behavior is a major challenge in biology.

Complex systems are able to adapt to changes in their environment. Human cells display adapted features obtained during three billion years of evolution; for instance, a core set of hundreds of genes is conserved across most known species [19]. In malignancies, cancer cells have a remarkable ability to develop

drug resistance using accelerated genetic and epigenetic adaptation [20, 21]. This is a major challenge in cancer treatment.

2.2 Quantifying complexity dimensions

There is no single measure for complexity that is applicable in all situations [22], but for the purposes of this thesis, we use well-established techniques from computer science for characterizing the two dimensions of complexity: quantity (volume) and quality (structure) [23]. Our approach is conceptual: the techniques illustrate strategies for understanding complexity, but the goal is not to derive specific equations or quantities.

Complexity dimension: volume

The first complexity dimension represents the amount of resources required by an object or a process. Examples include the number of nucleotides in a genome, or the time and memory required by an analysis program to process genomic measurements. Analysis of the amount of resources can be done using an asymptotic approach that is used in the analysis of algorithms [24, 25]. In this approach, we analyze the system as its size grows indefinitely. The basic question is: given a computational problem, such as sorting n numeric genomic coordinates, how many elementary computer steps are required, in the worst case, when n grows without bounds, i.e., $n \to \infty$?

To address this question, we select an algorithm that solves this problem and deduce from its description how rapidly the number of required steps grows. The sorting problem can be solved, for instance, by a merge sort algorithm [24] using a number of steps that grows with a rate of the order of $n \log n$, denoted $O(n \log n)$. There are other, less efficient algorithms for this problem that grow with the rate of n^2.

This mathematical framework gives us classes of growth rates that are useful for characterizing the time and space requirements of a process. An $O(n)$ (linear) process is less complex than an $O(n^2)$ (quadratic) process, which would be preferable to an $O(2^n)$ (exponential) process. As a practical rule of thumb, processes with polynomial complexity, i.e., $O(n^k)$ for constant k, are considered computationally tractable [25]. For harder complexity classes, exact solutions can be computed only for small data sets.

Similarly to measuring time, space complexity can also be analyzed using the $O(\cdot)$ formalism. In biology, the number of distinct proteins encoded by a genome having n genes, in the absence of alternative splicing, is $O(n)$. In contrast, the number of PPIs is, in the worst case, $O(n^2)$ – a more complex process. Empirically, however, the number of PPIs is closer to $O(n\sqrt{n})$ [26], and alternative splicing may increase the number of proteins beyond $O(n)$ [27].

Complexity dimension: structure

The second complexity dimension relates to the structure and information content of the system. Conceptually, this is a more challenging dimension to understand and there are multiple potential quantifying approaches [1]. The simplest approach, used in algorithmic information theory, is Kolmogorov

complexity (KC): the complexity of an object x is the length $K(x)$ of the smallest computer program that produces x when executed [28, 1]. For example, the KC of the human genome is the length of the shortest program that prints the genome.

Kolmogorov complexity has the technical limitation that its exact value is uncomputable in the general case. That is, it is possible to establish an upper bound for $K(x)$ by constructing a program that reaches this bound, but it is not possible, in general, to be certain that this is the best solution. For the human genome, an upper bound for KC is approx. 600 megabytes [29]. Repetitive regions, which comprise half of the genome, have lower KC than other regions.

2.3 Managing complexity

With an understanding of properties and dimensions of complex systems, we now ask, how can we manage complexity in biomedical research and development of methodology for high-throughput cancer experiments? We do so by observing three common organizational properties for managing complexity in both evolved (biological) and designed (engineered) systems.

Complexity management: parallelism

The first property is *parallelism*: complex systems often have a large number of actors that work concurrently. Biological examples include individuals, cells and proteins. Parallelism is used in high-throughput measurement devices to increase cost-efficiency, and in computing systems to decrease run time [30]. Parallelism helps solve challenges in the resource complexity dimension.

Complexity management: modularity

The second property is *modularity*: complex systems are divided into distinct parts that have high internal cohesion and limited coupling to other parts [31, 32]. For example, a cell is chemically isolated from its environment by the semipermeable plasma membrane [2]. In software engineering, modularity allows substituting one software module with another without large-scale refactoring to the rest of the system [32]. An important instance of modularity is hierarchy, in which lower-level entities are contained in a higher-level entity. Life is organized as a hierarchy with levels including: ecosystem $\rightarrow$ species $\rightarrow$ population $\rightarrow$ individual $\rightarrow$ organ $\rightarrow$ cell $\rightarrow$ organelle $\rightarrow$ molecule [33]. Another example is the structure of proteins, which comprises the following levels: primary (amino acid chain), secondary (alpha helixes and beta sheets), tertiary (three-dimensional folding) and quaternary (protein complexes) [2]. Composition and hierarchies also make complex systems more understandable for humans, and thus aid in managing the structural complexity dimension.

Complexity management: abstraction

The third property is *abstraction*, in which irrelevant details of an object are removed and the object is described using a simpler model [34]. Specifically, when an object possesses a given set of properties, abstraction is the omission of chosen properties. In molecular biology, an example is cell communication, in which cells present an interface of plasma membrane receptors and send messages to other cells using messenger molecules [2]. The internal cellular

logic is hidden (omitted) from the outside. This property is a key tool in both understanding biological systems and engineering complex software. Closely related to abstraction, *idealization* is the process of misrepresenting some properties of the object in order to simplify the model [34]. For example, a gene may in general prevent cancer from occurring, but under certain circumstances, it can instead promote cancer progression. Describing the gene as having only anti-cancer properties is an idealization (see Sec. 3.2.3 for a concrete example).

3 Cancer biology

Complexity dimension: structure

Cancer is a group of diseases characterized by abnormal cellular growth and malignant tissue invasion [4, 35]. It is among the most lethal diseases, leading to deaths of 7.6 million people yearly worldwide, and this figure is expected to rise in the future [5]. Cancer is also a highly complex disease [16]: tumors develop gradually over years or decades, and harbor mutations in numerous genes. Further, cancer is heterogeneous at three levels, which increases complexity: First, there are more than a hundred cancer types, grouped by tissue of origin, each having their distinct molecular characteristics. Second, patients of a given cancer have individually evolved tumors. Third, cells within a single tumor are heterogeneous at the genetic level [36]. These factors provide challenges for systematic characterization of cancer mechanisms.

Cancer incidence and mortality rates differ significantly depending on cancer type and gender. Worldwide and across genders, the most deadly cancers are lung (1.4 M deaths), stomach (0.74 M deaths) and liver (0.70 M deaths) cancers [5]. For men, prostate cancer has the second highest incidence (0.90 M cases), but relatively low mortality (0.26 M deaths). For women, breast cancer has both the highest incidence (1.38 M cases) and mortality (0.46 M deaths). Survival rates, defined as median time until cancer-related death, differ from 12 months in glioblastoma [37] to more than 15 years in prostate cancer [38]. Thus, the heterogeneity of cancer is also manifested in outcomes.

As a genetic disease, cancer has two main causative factors: inherited cancer risk alleles, and environmentally caused somatic mutations. Although there are much-publicized inherited cancer risk genes such as *BRCA1* and *BRCA2* [39], somatically acquired mutations are the main factor, accounting for the majority of cancers [4]. Well-known carcinogens, i.e., cancer-inducing physical agents, include ionizing radiation, tobacco smoke and viral infections [5, 4]. Exposure to these agents is related to life style and is thus partially avoidable. However, cancer can arise without exposure to such external agents, as there are also endogenous carcinogens. Examples of such agents are the hydrogen and hydroxyl ions present in the water solvent inside the nucleus, inducing sporadic DNA damage [4].

3.1 Glioblastoma, breast and prostate cancer

In this thesis, we focus on three cancers: glioblastoma multiforme (GBM), breast cancer, and prostate cancer. In the landscape of all cancers, these cancers cover a wide range of molecular and clinical characteristics, and thus demonstrate many key properties of cancers.

Glioblastoma multiforme GBM is the most common and most severe brain cancer, defined as grade IV astrocytic glioma [37]. Despite intensive treatment regimens including surgery, radiation and drug therapy, median survival is at 12

months. Histologically, GBM is characterized by highly invasive and diffuse tumor mass that frequently undergoes necrosis, i.e., cell death under abnormal conditions. GBM displays heterogeneity both within a tumor and between patients. GBM tumors are classified into two groups based on their mechanism of formation [37]. Primary GBM, which comprise 90% of cases, occur *de novo* with no previous diagnosis of lower grade gliomas. In contrast, secondary GBM, with 10% of cases, have a history of lower grade tumors. The latter is more frequently seen in younger patients (under 45 years). These two types are clinically similar but differ at the genetic level.

Breast cancer Breast cancer is the most diagnosed cancer and responsible for most cancer deaths in females [5]. The most common form of breast cancer is invasive breast cancer (IBC), which accounts for 22% of all female cancers [40]. Histologically, IBC is an adenocarcinoma of the mammary epithelium; 40–50% of tumors originate in the upper outer quadrant of the breast [40]. Breast cancer is characterized by multiple subtypes with different survival properties. A commonly used, and clinically useful, classification of breast cancer is based on the expression statuses of human epidermal growth factor receptor 2 (HER2), estrogen receptor (ER) and progesterone receptor (PR) proteins, all of which increase cell proliferation and cancer progression [41, 21]. Tumors with a negative status for all three markers, triple-negative breast cancer, have a 77% five year survival rate, compared to 93% for other subtypes [41].

Prostate cancer Prostate cancer is an adenocarcinoma of the male prostate gland and is the second most diagnosed cancer in males, although only the sixth most lethal [5]. Epidemiologically, prostate cancer is associated with old age, with more than 75% of men of age 85 developing this disease [42]. However, most cases are non-metastatic and relatively harmless; a key goal in prostate cancer treatment is targeting those tumors that are potentially aggressive. Whereas breast cancer is associated with ER, most prostate cancers depend on androgen hormones via the androgen receptor (AR). Tumors that continue to proliferate after pharmacological castration, i.e., reduction of circulating androgens, have a poor prognosis of two to three years [43], compared to more than 15 years for other prostate cancers.

3.2 Molecular hallmarks

Complexity management: modularity

Complexity management: abstraction

At the molecular level, cancer involves hundreds of genes interacting in complex networks. Such mechanisms are better understood by modularization, i.e., dividing genes into distinct functional pathways, and abstraction, i.e., assigning a well-defined cellular function to each gene. Such a process was followed by Hanahan and Weinberg, resulting in the "hallmarks of cancer" [35, 3] (Figure 1). These hallmarks are cellular features that are thought to be required for malignant cancer to arise and persist. We introduce the features of each hallmark and also discuss one gene or gene family related to each hallmark,

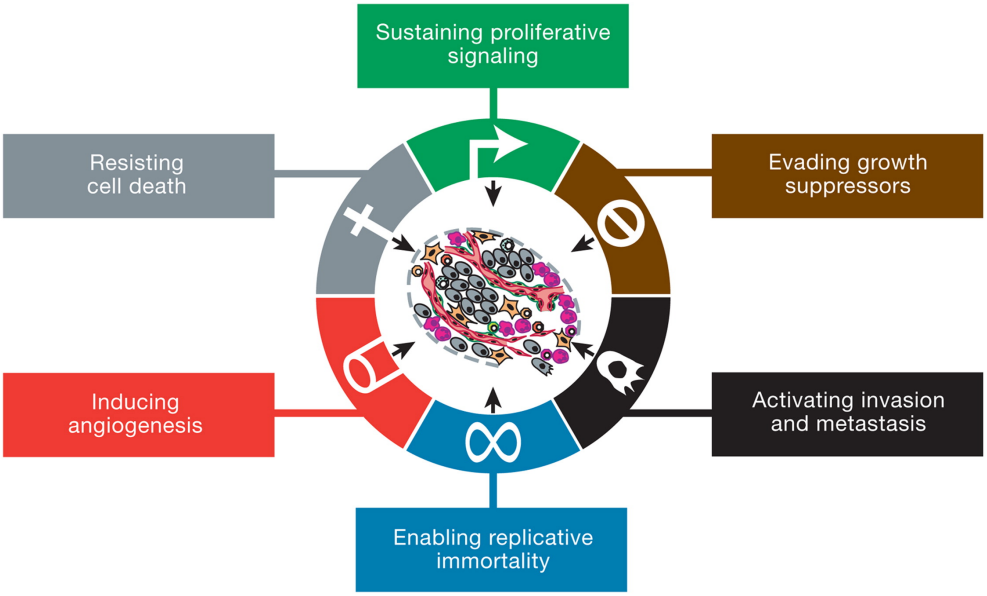


Figure 1: Hallmarks of cancer. Reprinted from [3] with permission from Elsevier, copyright 2011.

focusing on genes that are relevant for the three cancers studied in this thesis. Discussing specific instances of the hallmarks is an example of *concretization*, the complement of abstraction.

3.2.1 Proliferation

The most significant feature of cancer cells is their ability to grow and divide in an uncontrolled fashion [3]. In healthy tissue, the growth of cells is robustly regulated in order to maintain proper tissue structure. Mechanistically, this regulation occurs through signaling molecules and corresponding receptor proteins that relay the signal to the cells, usually by altering transcription. For example, an epithelial cell may be normally in a non-dividing state, but is induced by signaling to divide after wounding of nearby tissue [4]. In cancer, this regulatory link is broken, and the cell divides regardless of external signals, and is thus self-sustaining. This can occur through a variety of mechanisms, including overexpression of receptors leading to hypersensitivity to signals; self-production of signaling molecules for autocrine signaling; and mutation of receptors or other signaling proteins for ligand-independent activation [3]. Genes with potential for inducing uncontrolled division are members of *proto-oncogenes*; their mutated, cancer-inducing variants are *oncogenes*.

Nuclear receptors Both breast and prostate cancer are dependent on female and male sex steroids, estrogens and androgens, respectively. These steroids establish their effects through the nuclear receptors ER [44, 45] and AR [46, 43]. For estrogen, there are two known receptors, ERα and ERβ, of which ERα is better characterized and is thus focused on here.

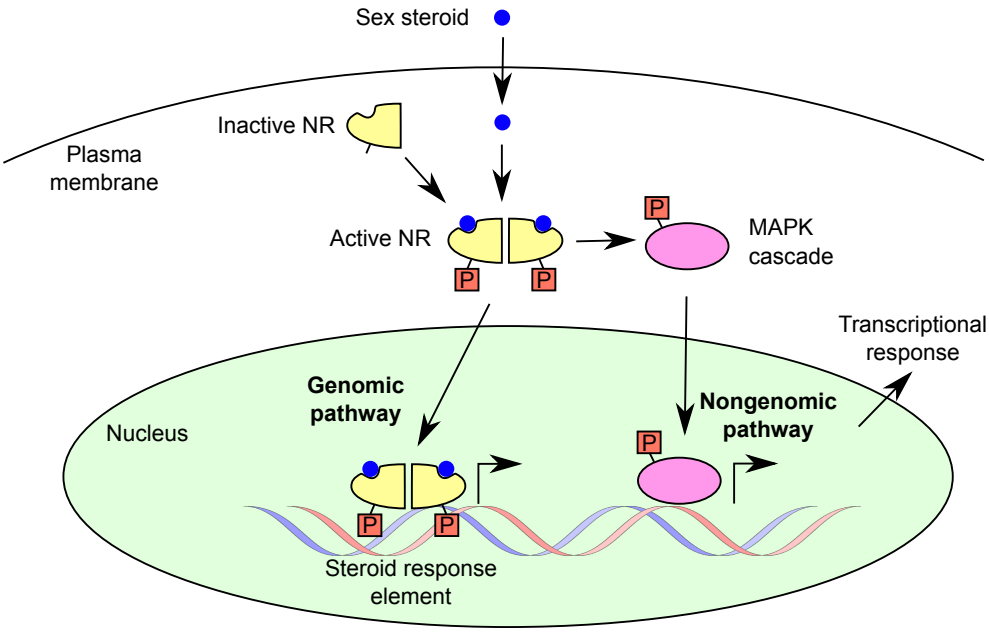


Figure 2: Genomic and nongenomic pathways of the nuclear receptors ERα and AR. MAPK, mitogen-activated protein kinase; NR, nuclear receptor; P, phosphorylation.

As transcription factors, ERα and AR affect cells in a multitude of ways; we have here selected their induction of proliferation as the representative hallmark. Nuclear receptors reside in the cytoplasm in their passive form, and upon binding to their cognate steroid ligands, they are phosphorylated, dimerized and transported to the nucleus, where they bind to DNA and alter gene expression [47] (Figure 2). In addition to this genomic pathway, nuclear receptors function in a nongenomic fashion by affecting signal transduction pathways, such as the mitogen-activated protein kinase (MAPK) pathway, in the cytoplasm [47, 45]. Sex steroid signaling is a common therapeutic target in breast and prostate cancer: treatments include inhibition of receptor function using drugs such as tamoxifen [48, 21] and bicalutamide [43]. Some tumors, however, are able to continue using nuclear receptor signaling by over-expressing the receptors or mutating them to become ligand-independent. Although ERα and AR are usually considered as "female" and "male" steroid receptors, breast cancers do respond to androgens and prostate cancers to estrogens, although using different mechanisms compared to their "native" hormones [49].

3.2.2 Evading growth suppressors

To counter the malignancy risk posed by proto-oncogenes, human cells have *tumor suppressor genes* (TSGs) that limit growth and division [3, 4]. Their presence explains why cancer generally occurs during old age and for only a subset of the population, despite the cancer risk individually held by 10^{13} cells in the body [50]. These genes integrate external signals, as well as monitor

internal cell states, and are able to halt cell cycle progression when necessary. Among the internal states monitored is the presence of DNA damage, which must be repaired before cell division can commence. To become malignant, cancer cells must inactivate key growth-limiting TSGs, usually with genetic or epigenetic silencing [3].

Retinoblastoma In GBM and other cancers, retinoblastoma (RB1) is a key negative regulator of the cell cycle and thus a TSG that often undergoes inactivation [51, 37, 4]. RB1 controls the restriction (R) point of the cell cycle, a critical decision point that determines whether the cell advances from the growth phase G_1 to DNA synthesis phase S and subsequently to mitosis. RB1 functions by binding to a class of mitogenic E2 transcription factors (E2Fs), repressing their function. After external mitogenic signals have accumulated, RB1 is phosphorylated and disassociates from E2Fs, which proceed to advance the cell cycle past the R point. RB1 can be deactivated by genomic deletion, promoter DNA methylation, or by alterations in the RB1 pathway that controls the phosphorylation state of RB1. In GBM, *RB1* is mutated in 25% of the cases, and more than 50% of the tumors have inactivated *CDKN2A*, a TSG on the RB1 pathway [37].

3.2.3 Evading apoptosis

Some tumor suppressor genes are able to induce programmed cell death, or *apoptosis* [3]. This cellular defense mechanism sacrifices the cell in a controlled fashion in order to prevent aberrant cells to affect the rest of the individual. Conditions for apoptosis include unrepairable DNA damage and lack of oxygen (hypoxia). Cancer cells evade apoptosis by silencing apoptosis-inducing genes or activating apoptosis-repressing genes [3].

p53 The *p53* gene, dubbed the "guardian of the genome" [18], is possibly the most important cancer-associated gene. It is mutated in half of all tumors [18, 52] and indirectly inactivated in most cancers. It is directly mutated in 38% of GBM [53], 37% of breast cancers [54] and 42% of prostate cancers [55].

p53 is a TF that monitors the health of a cell and its genome, and upon detecting abnormal conditions, halts cell cycle progression, activates DNA repair machinery, or induces apoptosis [18, 3]. In tumors, p53 is inactivated by missense (amino acid changing) mutations that change the function of the protein [4]. In the nucleus, p53 forms homotetramers, whose function is disrupted by the mutant form. In this way, one mutant allele displays dominant-negative behavior which disables the functionality of 15/16 tetramer configurations. The remaining 1/16 configurations are disabled by loss of heterozygosity, i.e., the loss of the remaining functioning allele. To complicate the interpretation of p53 mutations, some mutants of *p53* gain oncogenic functionality [56]. Under normal conditions, p53 is continuously transcribed and degraded at a similar and fast rate. When conditions such as DNA damage, hypoxia or abnormal levels of

mitotic proteins are detected, p53 is protected from degradation by phosphorylation, causing its protein levels to rapidly rise, which leads to modulation of transcription of the appropriate response genes.

3.2.4 Telomere elongation

Another functionality cancer cells must acquire is elongation of telomeres, the single-stranded DNA fragments at the ends of each chromosome composed of a repeating 5'-TTAGGG-3' sequence in humans [57]. With each cell division, telomeres are shortened by 30–150 nucleotides because the DNA polymerase active in mitosis can not fully replicate telomeric DNA [58]. When telomere length drops below a certain limit, cells enter into senescence, a non-replicative cellular state, or undergo death if replication is attempted. The human genome contains an enzyme, telomerase, that is capable of elongating telomeres [58]. However, this enzyme is not normally expressed in adult cells in humans, in contrast to, e.g., mice. Thus, cancer cells either activate telomerase or use a less-well understood alternative (ALT) pathway of telomere elongation [59].

Telomerase reverse transcriptase All three cancers studied in this thesis predominantly use the more common telomerase pathway for telomere elongation [60]. The telomerase enzyme is a complex composed of multiple units. One of these units, the human telomerase reverse transcriptase (TERT), is normally missing in healthy adult cells [58]. TERT, as the name implies, is a reverse transcriptase that transcribes a complement strand of telomeric DNA from an RNA template, which is housed within the complex. A regular DNA polymerase then synthesizes the actual telomere strand. Expression of *TERT* is re-activated by cancer cells, leading to immortalization [3].

3.2.5 Angiogenesis

Cancer cells interact with their tissue microenvironment. One crucial interaction is angiogenesis, the generation of new blood vessels, which provides the tumor with oxygen and nutrients, and disposes waste products [3, 4]. Angiogenesis is induced by secreting specific signaling molecules into the stroma. Neovascularization then occurs from existing nearby vasculature. The resulting vessels differ from those in healthy tissue: they are leaky and irregularly organized, but nevertheless functional [3].

VEGF In GBM, breast and prostate cancer, angiogenesis is frequently induced using the expression and secretion of vascular endothelial growth factors (VEGFs) within the tumor [37]. These signaling proteins are involved in *heterotypic interactions*: signaling between cells of different types [3]. VEGF originating from tumor cells bind to their corresponding receptors, VEGF receptors 1 and 2, in endothelial cells [61]. This promotes proliferation of endothelial cells in the direction of the VEGF signal, thus extending vasculature towards the

tumor. In healthy tissue, the growth of functioning vasculature involves several other signals as well, some of which are usually missing in tumors.

3.2.6 Invasion

A distinctive feature of malignant tumors, compared to premalignant lesions, is invasion of nearby tissue and metastasis to other tissues [3]. Cancer derives its lethality from this hallmark. Depending on cancer type, either invasion or metastasis is more important for disease outcome. For example, glioblastoma is highly invasive but rarely metastatic [37], whereas in breast and prostate cancer, death is associated with metastatic disease. Invasion and metastasis are complex processes that involve increased cell motility. A common metastasis model is as follows [3, 4, 62]: First, cellular motility is increased at the primary tumor, termed epithelial-mesenchymal transition (EMT). Next, mobile cancer cells enter blood or lymphatic vessels, and travel to another tissue. Then, they exit the vessels and form a new colony at the distant tissue. The EMT may be reversed via a mesenchymal-epithelial transition (MET). The cells may ultimately form a macroscopic tumor if they are viable in the new environment. Prediction of metastatic potential of the primary tumor is still challenging [62].

E-cadherin In several carcinomas, including breast and prostate cancer, E-cadherin is an important protein involved in invasion of tissue [3, 4]. E-cadherin is a transmembrane homodimerized protein that forms *adherens junctions* [2] with neighboring cells in epithelial tissue. These junctions are formed by coupling the actin cytoskeletons of cells together; E-cadherin binds to the cytoskeleton in its host cell and to other E-cadherins in neighboring cells. The loss of E-cadherin leads to loosening of cell-cell contact, increased motility and invasion. Tumor cells inactivate E-cadherin by downregulating the expression of the *CDH1* gene encoding this protein, or by mutating the gene [3]. As a non-carcinoma, GBM generally does not obtain invasiveness through E-cadherin inactivation; rather, it uses other mechanisms, such as dysregulation of matrix metalloproteinases [37].

3.3 Genetic mutability

The most important mechanism for cancer cells to obtain the molecular hallmarks is genetic mutability [3]. Healthy cells maintain highly stable genomes; the DNA is the most stable component of a cell. This is achieved using multiple damage protection mechanisms at anatomical, structural and enzymatic levels. In cancer, some of these mechanisms are damaged, and cells obtain a malignant phenotype. Thus, understanding DNA protection mechanisms is helpful in understanding carcinogenic processes.

The damage protection mechanism that cancer actively circumvents is the enzymatic mechanism, which is established by numerous enzymes that monitor and repair DNA damage [63]. For instance, DNA polymerases, which replicate DNA

during the S phase of the cell cycle, have proof-reading and error correction capacities that correct many replication errors. As a result, DNA replication has an error rate of one in 10^{10} bases in human [64]. Likewise, there are enzymes that detect abnormal chemical structures in chromatin. These mechanisms are often deactivated in cancer, resulting in greatly increased mutation rates.

The resulting mutations include both small-scale and large-scale DNA changes. The former include single nucleotide variations (SNVs), and small insertions and deletions (indels) [52]. The latter include changes in overall chromosome numbers (aneuploidy) as well as changes in chromosomal structure, such as translocations between chromosomes or large-scale amplification of chromatin [65].

3.4 Gene regulation

The effects of genetic changes in cancer cells are translated to cellular behavior through several mechanisms, one of which is aberrant gene expression, i.e., over- or under-expression of genes. Gene expression in healthy cells is a highly regulated process; multiple regulatory mechanisms control when, where and how much gene product (protein or RNA) is produced from the gene DNA template [66, 2]. In fact, a substantial portion of the human genome is thought to be involved in gene regulation [67], explaining the genomic complexity of higher eukaryotes despite having a similar number of genes as single-celled eukaryotes.

Gene expression is a multi-step process that is initiated by transcription factors (TFs), DNA-binding or associated proteins that modify chromatin conformation and recruit other TFs to chromatin [66] (Figure 3). TFs can function in a combinatorial fashion, so that binding of one TF is required for other TFs to bind. Alternatively, a particular TF may prevent others from binding. TFs bind either to promoter regions of genes – segments immediately upstream of transcription start sites (TSSs) – or to enhancer elements further away from TSSs. TFs recruit to chromatin a protein complex, RNA polymerase, that catalyzes the transcription of template DNA to RNA. There are three nuclear RNA polymerases in human. We here focus on RNA polymerase II (PolII), which transcribes most protein-coding genes [2]. To complement PolII, RNA polymerase I transcribes ribosomal RNA, and RNA polymerase III transcribes other non-coding RNA, as well as a subunit of ribosomal RNA.

PolII synthesizes a single-stranded RNA molecule that is then processed by spliceosomes, RNA-protein enzymes that cut out introns and selected exons [68, 2]. This process gives rise to alternative splicing, i.e., the production of multiple proteins from one gene. The RNA molecule is matured into messenger RNA (mRNA) by attaching a polyadenosine tail to its 3' end and capping its 5' end with a modified guanosine. The resulting mRNA molecule is exported from the nucleus and translated into a protein by ribosomes. Proteins are subject to post-translational modifications (PTMs), such as covalent attachment of phosphate

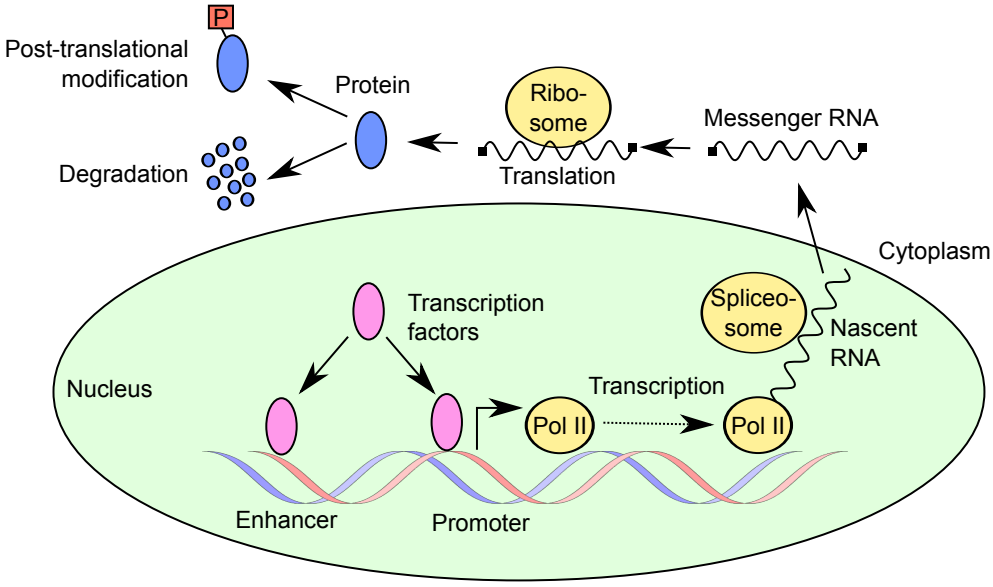


Figure 3: The mechanics of gene expression. The process is initiated by transcription factors binding to DNA, which recruits a PolII complex to the transcription start site. PolII synthesizes a nascent RNA molecule which is subject to splicing by spliceosomes. After both ends of the RNA molecule are modified, the mature messenger RNA is exported to the cytoplasm and is translated to protein by ribosomes. Proteins may undergo post-translational modification, such as phosphorylation (P), and are ultimately degraded.

groups to serine, threonine or tyrosine residues [69, 2].

Regulation of gene expression occurs at multiple phases. TFs are regulated by their production – whether a given TF is present in a cell – and their molecular state – for example, whether a particular amino acid is phosphorylated. Cellular localization is also a TF regulation mechanism: some TFs, such as estrogen receptor, can reside both in the cytoplasm and in the nucleus [45]. Splicing and thus the production of specific protein isoforms is a regulated process, as evidenced by different isoform distributions in different tissues [70]. For proteins, the major regulation mechanisms are PTM and, ultimately, degradation.

Epigenetics plays an important role in gene regulation [71]. Epigenetics refers to non-genetic mechanisms that affect gene expression without affecting the primary DNA sequence. Two major mechanisms of such are DNA methylation and histone alterations. DNA methylation is the attachment of methyl groups to carbon 5 of cytosines [72]. Methylation of cytosines in promoters is associated with suppression of gene expression, whereas methylation in other parts of the genome is less well characterized [73]. There is evidence that methylation regulates alternative splicing [74]. Cancer genomes are aberrantly methylated; globally, they are hypomethylated, i.e., lack methylated cytosines compared to healthy cells. However, specific genes, such as tumor suppressors, are methylated in promoters in order to silence their expression.

Histone alterations are PTMs or substitutions of histones [75]. Histones are

building blocks of nucleosomes, low-level organizational units of chromatin composed of hetero-octameric histones. Their N-terminal tails are subject to PTMs such as methylation and acetylation, and the state of modifications affects transcriptional activity. An example of a histone modification is trimethylation of lysine 4 of histone H3 (H3K4me3), which is associated with the promoters of actively transcribed genes [75]. Another type of histone alteration is the substitution of a histone in the nucleosome octamer. An example of such substitution is H2A.Z, which can replace histone H2A [76]. Like H3K4me3, this alteration is also associated with active transcription.

4 High-throughput measurement techniques

Complexity management: parallelism

Given the complexity of the human genome and the variety of genotypic and phenotypic alterations in cancer, a challenge in cancer research is efficiently measuring and characterizing these alterations. In principle, this could be done using low-throughput measurement techniques that query one or a few genomic loci at a time. However, this is infeasible in practice due to the high cost. Rather, genome-wide experiments are conducted using high-throughput measurement techniques that use parallelization at the experimental level to increase efficiency and lower cost.

Parallelization at the experimental level mirrors the underlying parallelism present in the subject of measurement, i.e., human cells. In a common experiment, numerous cells of similar phenotype are analyzed together, and each cell contains a genome comprising 3 billion base pairs and 20,000 protein-coding genes. This biological parallelism can be utilized at the experimental level by first dividing cellular samples into aliquots that are analyzed in parallel. Then, within the aliquots, genomic regions or genes can be measured in parallel. Prominent technologies using this two-step parallelization are DNA microarrays and deep sequencing.

4.1 DNA microarrays

DNA microarrays are a popular high-throughput technology for measuring genome-wide gene expression, genomic variation, and DNA–protein interactions [9, 77, 78] (Table 1). Inspired by protein arrays in late 1980s [9], DNA microarrays were introduced in 1995 [79]. After the technology matured in late 1990s and early 2000s, microarrays became a standard tool for genome-wide transcriptomics, single nucleotide polymorphism (SNP) genotyping, and copy number investigation [80]. In recent years, sequencing-based approaches (Section 4.2) have generated comparatively more interest, but microarrays remain a useful tool, in particular for expression profiling [81]. However, in some areas, such as DNA–protein profiling, sequencing has replaced microarrays.

Table 1: DNA microarray applications.

Technology	References	Purpose
Expression arrays	[81]	Quantifying gene expression (mRNA or microRNA); splice variants.
SNP arrays	[80]	Detecting SNPs in a population; copy number changes.
Array CGH	[80]	Large-scale copy number changes.
Methylation arrays	[80]	DNA methylation.
ChIP-chip	[80, 82]	DNA–protein interactions.
Sequence enrichment	[83]	Selective deep sequencing of, e.g., whole exomes.

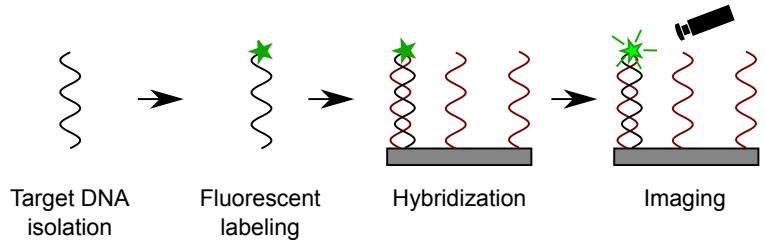


Figure 4: DNA microarray protocol. First, target DNA is isolated; in gene expression profiling, this also includes a reverse transcription step to obtain DNA from RNA. Next, target DNA molecules are labeled with a fluorescent dye. Labeled DNA molecules are hybridized with probes fixed on a microarray. Finally, the dyes are excited using a laser and the fluorescence is imaged.

DNA microarrays are based on fluorescently labeled target DNA hybridizing with an array of DNA probes attached to a surface, allowing quantification of target DNA matching each probe [9] (Figure 4). The array surface is a glass slide or a silicon chip, depending on the manufacturer. Probes in commercial microarrays are short (25 bases) or medium length (50–70 bases) single stranded oligonucleotides; long complementary DNA (cDNA) molecules can be used in home-printed microarrays. Short and medium length probes are synthesized *in situ* on the surface using photolithography or ink-jet printing. To increase fluorescent signal strength, probes are arranged as clusters of identical sequence, forming *spots* that are analogous to DNA colonies in second-generation sequencing. Spots are 10–100 μm in diameter and over a million spots fit on one array [77].

The experimental protocol for microarrays includes reverse-transcribing target RNA to DNA (in expression studies), labeling the resulting molecules with a fluorescent dye such as Cy3 or Cy5 [84], and allowing the labeled molecules to hybridize (anneal) with probes on the array. After hybridization, the fluorescent dye (or dyes for multi-color protocols) is excited with laser, and the array is imaged. The quantity of emitted photons from a certain spot correlates with the amount of target DNA whose complementary sequence matches the probe. The resulting images are quantified and the numerical values are analyzed statistically.

Probe design is a critical step of microarray manufacturing and is guided by the intended application for the microarray [9, 80]. For expression profiling, probes are selected from the sequences present in RNA molecules, such as 3' untranslated regions (UTRs) and exons. Depending on the design, either a single or multiple probes target one gene. High-density Affymetrix Human Exon arrays contain probes for individual exons, thus allowing quantification of splice variants [85]. A problem inherent with the *a priori* method of probe design is that some probes on commercial microarrays may be invalid or difficult to interpret. For example, a microarray designed based on an earlier version of a reference genome may have probes that do not align to a newer version of the reference. For the Agilent 4x44K expression array, 40% of the probes have

been found to have design defects [86].

Microarrays have a two-way interaction with sequencing technology. First, probe design for commercial microarrays requires a reference genome for the species and, for expression microarrays, identification of gene locations. Thus, the sequencing of the human reference genome greatly increased the usefulness of microarrays for human studies. Second, microarrays can be used to select specific genomic regions for targeted deep sequencing, such as exomic regions [83]. This reduces costs and allows increasing sequencing coverage in these regions, leading to increased sensitivity for detecting genomic variants.

4.2 Deep sequencing

Deep sequencing technologies enable determination of nucleotide sequences in a genome-wide fashion. Deep sequencing has numerous use cases from sequencing genomes of model organisms to quantifying RNA expression levels.

4.2.1 History of DNA sequencing

The need for determining sequences of biological polymers was identified in the late 1940s, when it was observed that amino acids in peptides are arranged in a linear sequence that is neither fully arbitrary nor periodic [87]. A critical event for enabling DNA sequencing was the discovery of the molecular structure of DNA in 1953, which was accompanied with a realization of how DNA is replicated [88]. The "Watson–Crick" pairing of nucleotides in DNA replication is used by all current DNA sequencing technologies. The first nucleotide polymer to be sequenced was transfer RNA of *E. coli* in 1965 [87]; DNA sequencing began maturing in the 1970s.

The first generation of DNA sequencing, Sanger sequencing, was developed in 1977 as an improvement of related earlier techniques [89]. Sanger sequencing uses a DNA polymerase from *E. coli* to synthesize a second strand for the DNA target molecule, and specifically engineered nucleotides that terminate the synthesis. These dideoxynucleotides (ddNTP) lack hydroxyl groups at both the 2' (as in regular DNA) and 3' (only in engineered nucleotides) carbons and thus halt further DNA replication. By using four pools of nucleotides composed of regular and engineered nucleotides, radiolabeled phosphates, and gel electrophoresis to separate polymers based on length, the original Sanger sequencing enabled sequencing reads of up to 200 nucleotides [89].

Complexity management: parallelism

By mid 1990s, Sanger sequencing had been improved and automated by replacing radiolabeling with fluorescent markers and gels with capillaries, and by using 96 well plates for limited parallelization [87]. This enabled establishing sequencing "factories" that could sequence millions of bases daily [87]. The read length increased to 1000 bases and base accuracy was high (99.999%) [90]. Growing data volumes and the need for their interpretation motivated the establishment of sequence data centers and bioinformatics development efforts. The

use of Sanger technology culminated in the sequencing of the human genome, first published in 2001 [91, 92]. Despite its high cost ($3.8 billion), the human genome project has provided a 65 fold return on investment [93].

Complexity management: parallelism

Despite the usefulness of Sanger sequencing, it remains expensive and slow when considering the size of the human genome. In the mid-2000s, improved technologies appeared [8]; these are collectively called next-generation sequencing or second-generation sequencing (SGS). Their main advantage is massively increased parallelism and lower cost per base. These are obtained from three improvements to Sanger technology [90]. First, SGS technology uses imaging of arrays of DNA colonies in contrast to capillaries, enabling massively parallel sequencing of millions of colonies. Second, source DNA is amplified using specifically engineered DNA polymerase based protocols, instead of *in vivo* bacterial plasmid amplification used in shotgun Sanger sequencing. Third, reagent costs are reduced by using a common reaction volume (e.g., a glass slide), which distributes reagents to all colonies of the array. The main drawback of SGS is the generally reduced read lengths; this is ameliorated by the presence of reference genomes, which are used for aligning the short reads. Many commercial SGS technologies are available [8]; in this thesis, we focus exclusively on the Illumina/Solexa platform (Illumina Inc., San Diego, USA).

4.2.2 Illumina sequencing platform

Illumina is a major provider of second-generation sequencing platforms, with products such as HiSeq 2500 and HiSeq X for high-throughput sequencing, and MiSeq for benchtop sequencing [94, 95]. The technology was developed by the UK-based Solexa Ltd., which was sold to Illumina Inc. in 2006 [96]. Illumina has refined the technology since its introduction: current HiSeq 2500 systems can produce up to 600 Gb of sequence per run, compared to 4 Gb of the original Genome Analyzer [97].

Sample preparation and adapter ligation The general Illumina sequencing protocol [97] (Figure 5) starts with the preparation of a DNA library, whose details depend on the deep sequencing application in question. Library preparation yields two-stranded DNA fragments that are 200–2000 bp in size. To both ends of these fragments, a proprietary adaptor is ligated, producing blunt-ended sequencing templates.

Bridge amplification Templates need to be amplified because imaging sensors are not sensitive enough to detect single molecules. This is done using bridge amplification on a flow cell (glass slide with eight independent lanes), a protocol that produces spatially clustered colonies of DNA molecules having the same sequence [97]. First, templates are denatured and adaptors are annealed to their complementary sequences fixed on the flow cell, producing single-stranded templates attached to the surface. Then, a second strand is synthesized by DNA polymerase using the fixed adaptor as primer and the original strand is washed

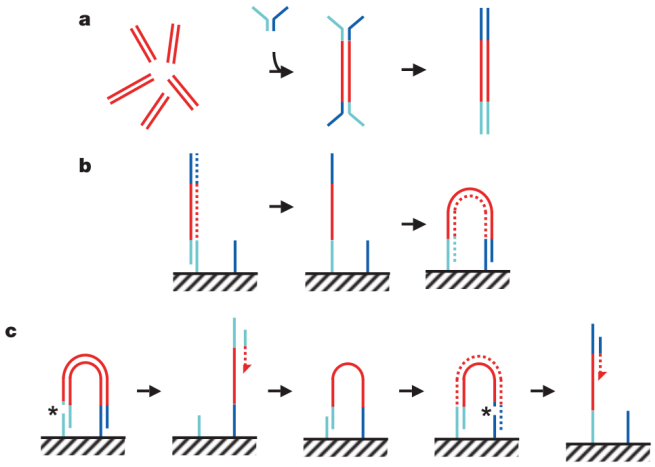


Figure 5: Illumina sequencing protocol. (a) DNA is fragmented and a sequencing library is generated by ligating adapters to both ends of the fragments. (b) In bridge amplification, a denatured template is annealed to a fixed adapter and the complement oligonucleotide of the template is synthesized. After washing away the original template, a bridge is formed and a new strand is synthesized. The process is repeated several times. (c) In sequencing by synthesis, DNA is first linearized by cleavage ($\star$) and the first read (dotted) is generated from the free end of the template. For paired-end sequencing, a bridge is formed and the second read is generated from the other end of the template. Adapted by permission from Macmillan Publishers Ltd from [97], copyright 2008.

away. In the "bridge" step that follows, the free end of the template is attached to the surface by annealing its adapter to a second fixed complementary oligonucleotide. This template, attached to the surface from both ends, is amplified to yield two templates fixed to the surface. This process is repeated, creating spatial clusters.

Sequencing by synthesis Sequencing is done using custom nucleotides called reversible terminators [97, 8]. These nucleotides have an azidomethyl group in the 3' carbon, blocking synthesis similar to ddNTP used in Sanger sequencing, and a fluorescent marker attached to the nucleobase. In contrast to Sanger sequencing, both of these groups can be chemically cleaved, thus producing non-blocking nucleotides. There are four distinct fluorescent markers, one of each base. Sequencing proceeds in cycles, each producing one base of sequence. At the beginning of a cycle, nucleotides with terminators are incorporated to the flow cell and a modified DNA polymerase extends the complementary strand of each template by one base. Then, extra nucleotides are washed and the flow cell is imaged by exciting each fluorescent dye in turn using laser; this produces the raw data for further processing. Dyes and terminators are cleaved and the cycle is repeated for 36 to 150 times.

Paired-end sequencing To extend both the sequencing depth (total number of reads) and physical coverage (portions of the genome queried), paired-end sequencing (PES) is often employed. Introduced for Sanger sequencing [87],

this technique involves sequencing both ends of a template, usually with a gap between the obtained sequences. This allows obtaining information from a longer genomic region than with single-end sequencing. On the Illumina platform, PES is done after synthesizing the primary short read by creating a bridge from the template and sequencing the resulting oligonucleotide [97].

Common bioinformatics workflow Computational processing of sequencing data is a major challenge for deep sequencing experiments and the methods vary depending on the application. However, there are some common bioinformatics steps in Illumina-based experiments. First, raw images are analyzed by identifying and quantifying fluorescent clusters; artifacts such as shifted image coordinates between cycles are accounted for [97]. Then, bases are called from the quantified images, taking into account cross-talk between dyes. In this important step, each base is assigned a quality score that represents the probability of an incorrect base call [98]. From this point, workflows diverge based on the application. A common step, however, is short read alignment to a reference genome. In this step, the called sequence is compared against all or the relevant locations of a reference genome and, ideally, assigned to a certain location. In practice, some reads do not align to the reference genome, some align to multiple locations, and some align with one or more mismatches. Longer and higher quality reads are easier to align, thus motivating the refinement of sequencing technology. Short read alignment is done with specialized heuristic software such as ELAND [97], Bowtie [99] or BWA [100], because exact alignment algorithms [101] and general-purpose heuristics, such as BLAST [102], are inefficient.

4.2.3 Sequencing application: ChIP-seq

In addition to reduced cost, the improvement of sequencing technology has allowed broadening the scope of its applications. Whereas in the Sanger era the focus was on determination of reference sequences for model organisms, today deep sequencing is used for several other purposes (Table 2).

One important application, and a focus in this thesis, is the genome-wide measurement of DNA–protein interactions and histone alterations using chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) [82]. This technology allows investigation of gene regulation mechanisms during the early stages of gene expression: transcription factor binding, histone response, and activation and progression of the PolII complex. ChIP-seq is thus suitable for detecting, for example, what is the immediate response of a cancer cell to an external proliferative signal. ChIP-seq was one of the earliest second-generation sequencing applications [113] due to its relatively modest requirement for sequencing depth; the technology is also partially based on an earlier ChIP on microarray (ChIP-chip) technology, aiding in the transition.

ChIP-seq is based on chromatin immunoprecipitation, the purification of DNA fragments bound to a certain protein [114]. In the ChIP protocol, DNA–protein

Table 2: DNA sequencing applications.

Technology	References	Purpose
DNA reference sequencing	[91, 92]	Assembly of a reference genome.
DNA resequencing	[103]	Genetic variation in a population; personal genomics; disease susceptibility.
Tumor DNA sequencing	[104, 105, 54]	Somatic mutations and chromosomal aberrations; putative causative factors.
RNA sequencing	[106]	Quantification of gene expression (mRNA or microRNA); novel transcripts; fusion genes.
ChIP sequencing	[82, 67]	DNA–protein interactions; histone alterations.
Global run-on sequencing	[107, 108]	Quantification of nascent RNA.
Bisulfite sequencing	[109]	DNA methylation.
HiC	[110]	DNA–DNA interactions (chromatin conformation).
ChIA-PET	[110]	Chromatin conformation related to a protein of interest.
Ribosome profiling	[111]	Translation profiling at single codon resolution.
Metagenomics	[112]	Identification of microbial species in an environment; phylogenetics.

bonds are first fixed using formaldehyde (CH_2O), a chemical agent capable of reversibly crosslinking both DNA and protein molecules. Following formaldehyde treatment, DNA is fragmented using sonication to yield oligonucleotides with a length range of 200–600 bp [82]. Using an antibody against the protein of interest, the DNA–protein–antibody complexes are immunoprecipitated using a centrifuge and the formaldehyde crosslinking is chemically reversed. Finally, the isolated DNA fragments are sequenced to obtain the primary data.

ChIP-seq derives flexibility from the use of antibodies, which allows querying various types of DNA-binding proteins. At the same time, proper selection and validation of the antibody is important for a successful experiment, as up to a third of commercially available antibodies are suboptimal for ChIP-seq, and antibodies from different batches may have different properties [82]. Another important experimental factor is the correct duration of formaldehyde treatment [114].

Like most experiments, ChIP-seq needs a control experiment to account for factors not related to the research question [82]. For ChIP-seq, an adequate control sample is particularly important because the technology is quantitative. Factors affecting the validity of results include biases in sequencing and alignment of particular DNA regions. For example, open chromatin is easier to sequence, and short reads in repetitive genomic regions are difficult to align to the reference genome [82]. Several options are available for a control sample. A commonly used method is the sequencing of input DNA, i.e., direct genomic sequencing without immunoprecipitation. This controls for sequencing, copy

number and alignment biases, but also requires sufficient sequencing depth as the whole genome is covered [82]. A second option is the use of a nonspecific (mock) antibody or immunoprecipitation without an antibody. A challenge with this approach is that limited DNA material is pulled down and results are stochastic [82].

Bioinformatics analysis of ChIP-seq data starts with aligned short reads and the methods used depend on the type of protein used in ChIP and on the research goals. For TFs, the usual goals are to determine their binding sites (BSs) and putative DNA response elements. BSs are visible as peaks in the coverage of short reads along the genome. Statistically, the aim is to find regions with significantly more reads in the ChIP sample than in the control sample. There are numerous methods for detecting peaks [115]; a popular one is MACS [116]. Peak discovery can be followed by identification of DNA sequence motifs enriched at BSs [117, 118], or mapping peaks to genes [119], followed by gene-level analysis.

In contrast to TFs, ChIP-seq profiling of PolII usually has different goals and methodology. Although the goal may be to identify novel PolII binding sites [120], a more common goal is to identify actively transcribed genes and to quantify changes in transcription rates between samples [121]. ChIP-seq using a PolII antibody results in broad regions of short read enrichment, corresponding to actively elongating PolII complexes, as well as peaks at promoters. Methods tuned for TFs are not optimal to PolII [82]. For quantifying transcription, a common analysis method is to count short reads along each gene body and to compare these counts between samples.

5 Aims of the study

To goals of the thesis are to develop computational methods for analysis of genome-wide experimental data, and to apply these methods to cancer research. The specific goals are to:

1. Develop an umbrella software that manages the overall structure of genome-wide data analyses, is scalable in both complexity dimensions, and facilitates reuse of analysis code.

   Complexity dimensions: volume and structure

2. Develop methods for analyzing ChIP-seq data from experiments that profile PolII and histone alterations using a time series.
3. Develop modular and well-structured methods for preprocessing deep sequencing data.
4. Identify molecular aberrations in GBM, breast and prostate cancer and assess the significance of these aberrations to patient survival.

6 Material and methods

6.1 Biological material and methods

Biological material and methods are summarized in Tables 3 and 4, and full details are in the publications.

Table 3: Biological sample material.

Publication	Type	Samples
Publication I	Patient	338 primary GBM patients; public data set [53]
Publication I	Cell line	A172, U87MG, LN405 and SVGp12 (control) GBM cell lines; three replicates
Publication II	Cell line	MCF-7 breast cancer cell line [122]; time series
Publication II	Patient	150 primary breast cancer post-menopausal ER+/HER2– patients [54]; for validation, 130 [123] and 159 [124] primary breast cancer patients; all public data sets
Publication III & Related Publ. I	Cell line	LNCaP-1F5 [125] prostate cancer cell line
Related Publ. I	Cell line	VCaP prostate cancer cell line
Related Publ. I	Patient	350 prostate cancer patients with prostatectomy treatment

Table 4: Biological experimental methods. "Public" indicates that the experiments were done by authors of published articles and we (re)analyzed the data.

Publication	Samples	Method
Publication I	GBM patients	DNA microarrays: gene expression (mRNA, exon and miRNA), array CGH, SNP, DNA methylation (public)
Publication I	GBM cell lines	Gene silencing by siRNA for 11 genes
Publication I	GBM cell lines	Proliferation and apoptosis assays
Publication II	MCF-7	ChIP-seq for PolII, H3K4me3, H2A.Z and ER
Publication II	MCF-7	RNA-seq
Publication II	Breast cancer patients	Gene expression microarrays (public)
Related Publ. I	LNCaP-1F5 and VCaP	Gene silencing by siRNA for FOXA1
Related Publ. I	LNCaP-1F5 and VCaP	ChIP-seq for AR, FOXA1, GR (LNCaP-1F5 only) and IgG (control)
Related Publ. I	LNCaP-1F5 and VCaP	ChIP-qPCR for AR
Related Publ. I	LNCaP-1F5	ChIP-seq for H3K4me2
Related Publ. I	LNCaP-1F5	Deep sequencing for DNaseI hypersensitivity
Related Publ. I	LNCaP-1F5 and VCaP	Gene expression microarrays
Related Publ. I	Prostate cancer patients	Immunohistochemistry for AR for FOXA1

6.2 Scientific workflow management systems

Used in: **Publication I, Publication II, Publication III, Related Publication I**

High-throughput measurement techniques produce complex data sets that must be analyzed computationally. Construction and execution of analysis implementations can be done using several approaches that range from manual *ad hoc* interactive methods to fully automated structured methods. For the latter, two often used techniques are scripting in a programming environment, such as R and Python, and the use of workflow technologies. In this thesis, we use the workflow approach to organize the large-scale structures of analysis implementations.

A *workflow* is a structured collection of tasks to implement a business (e.g., scientific) process [126, 127] (Figure 6). This broad definition includes manual and semi-automated workflows, but here the focus is on fully automated workflows. The structure of a workflow defines dependencies, and thus valid execution orders, of tasks. A common representation of a workflow is a directed network [128].

Complexity management: parallelism

Complexity management: abstraction

Complexity management: modularity

Advantages of workflows include the following [127]. Workflows increase effectiveness through automation, as well as reuse of existing workflows and building blocks for tasks. The dependency structure of tasks allows automatically parallelizing workflow execution, speeding up execution and reducing manual parallelization effort [129]. Workflows aid in reproducibility of analyses through the formal definition of tasks and their relationships. Intermediate results of each step can be stored and inspected, which aids in diagnosing problems and allows using the cached results on repeated workflow executions. The formal structure of workflows is amendable for analysis and identification of common patterns [128, 130, 131], generating abstractions that facilitate understanding and communication of complex workflows, as well as comparison of workflow engines. One particularly useful pattern for structurally complex workflows is decomposition of the workflow into a hierarchy of sub-workflows, each of which is simpler to understand in separation.

In bioinformatics, workflows have been used for integrating heterogeneous data from public databases and WWW resources [127], and analyzing high-throughput data. Published scientific workflow management systems include graphical approaches, such as Galaxy [132], Taverna [133], GenePattern [134], Chipster [135], Kepler [136] and KNIME [137], as well as programming-based approaches, such as Biopipe [138], Pegasus [139] and Ruffus [140]. The differences between these two approaches are twofold. On the one hand, graphical approaches are easier to learn and use by non-programmers. On the other hand, programming-based approaches scale better to complex workflows. Differences between various workflow implementations include suitability for analysis of large data volumes, choice of programming language(s) or web service protocols for implementing elementary analysis steps, and availability of ready-made tools for specific experimental techniques, such as deep sequencing.

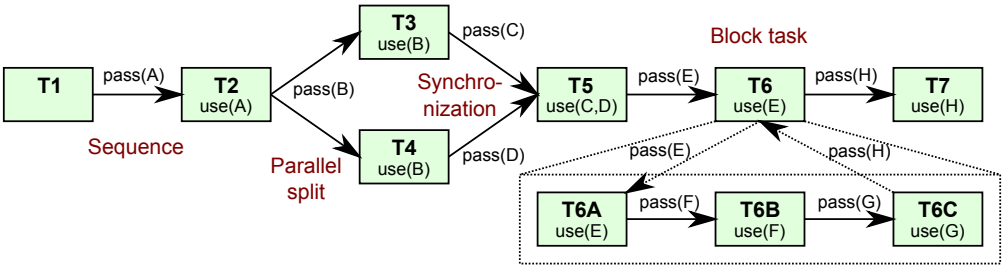


Figure 6: Example workflow with ten tasks, one of which (T6) is a composite task. The workflow demonstrates several workflow control and data patterns [128, 130]. Each task produces a certain output data, which is passed as input to the subsequent task(s): for example, the data item `B`, produced by task `T2`, is passed to tasks `T3` and `T4`. The simplest control pattern is *sequence*, which is a linear dependency between two tasks. A *parallel split* separates the workflow into independent flows, which are joined by a *synchronization* task (`T5`). A *block task* enables composition of the workflow into sub-workflows. The parent task `T6` passes input to the child task `T6A` and the result from the sub-workflow (`H`) is passed back to the main workflow.

6.3 Kaplan-Meier survival analysis

***Used in*:** **Publication I, Publication II, Related Publication I**

In fields such as medicine, there is often a need to estimate the time "survived" by a subject of study, such as a patient, until an *event*, such as death from a disease. Statistical methods for working with such use cases are called survival analysis methods. Two basic goals in survival analysis are the estimation of survival times (curves) of a population, and the comparison of survival times of two or more groups [141]. A challenge in survival analysis is the presence of incomplete data: a subject may be removed from the study before the event occurs because the follow-up time ends, the person withdraws, or dies from causes not related to the study. Collectively, these are called *censoring* the subject. Kaplan-Meier (KM) survival analysis is a popular method that estimates survival times by accounting for missing information [142]. Often coupled to KM is the log-rank statistical test used to compare survival times between subject groups.

Formally, the random variable $T > 0$ is the survival time of a subject before the event, and $\delta \in \{0,1\}$ indicates whether the subject was censored (0) or experienced the event (1) [141]. In case of censoring at time t, the precise survival time T is unknown; it is only known to be greater than t. Central to survival analysis is the survival function $S(t) = P(T > t)$, which is the probability that a subject survived beyond time t; this function is estimated by KM. Survival functions are decreasing, i.e., survivorship stays the same or drops over time, and their theoretical limits are $S(0) = 1$ and $S(\infty) = 0$. Empirical survival functions are step-wise functions with discontinuous drops at times of events.

KM obtains the maximum likelihood estimate $\hat{S}(t)$ of the survival function $S(t)$, i.e., it derives a statistical model that maximizes the probability of observed

data [142]. It does so by computing survival probabilities at each time of event separately, based on subjects observed at that time; probability of survival at time t is then the product of survival probabilities up to t [141]. That is,

$$\hat{S}(t_{(j)}) = \prod_{i=1}^{j} \mathrm{P}(T > t_{(i)} | T \geq t_{(i)}) = \hat{S}(t_{(j-1)}) \times \mathrm{P}(T > t_{(j)} | T \geq t_{(j)}), \quad (1)$$

where $t_{(j)}$ denotes the time of the j'th event in the ascendingly ordered list of event times. We define $\hat{S}(t_{(0)})$ as 1. The KM method is illustrated in Figure 7.

In addition to the survival point estimates $\hat{S}$, it is also important to estimate the variability of survival, i.e., confidence intervals. This is done separately for each time of event. The widths of confidence intervals are proportional to the variance of the point estimate $\hat{S}$ at each time point:

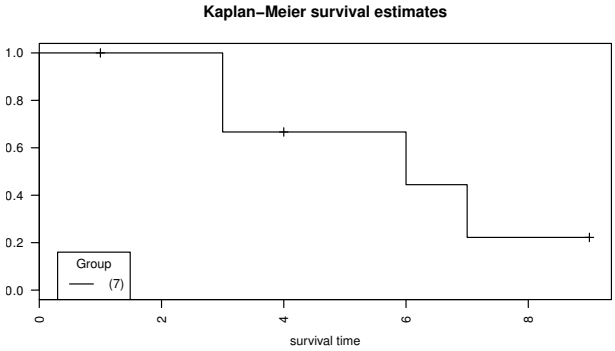
$$\mathrm{Var}(\hat{S}(t_{(j)})) = \hat{S}(t_{(j)})^2 \sum_{i=1}^{j} \frac{d_i}{n_i(n_i - d_i)}, \quad (2)$$

where d_i and n_i are the number of events and the number of remaining subjects at time $t_{(i)}$, respectively [142]. From this formula, we can see that confidence intervals intuitively become wider as the number of subjects n_i decreases. They also tend to widen in later time points due to summing of variances from earlier time points, although this effect is compensated by the decreasing survival estimate term $\hat{S}(t_{(j)})^2$.

The log-rank test is used for comparing survival times from two or more subject groups, such as treatment and control groups [141, 143], and thus it complements the KM method. It tests the null hypothesis that all KM curves are derived from the same distribution, i.e., there are no significant differences in survival. The log-rank test computes expected counts of events at each event time point based on aggregate events in all groups, and compares these to observed counts. With $k \in \{2, 3, \ldots\}$ subject groups, a statistic derived from expected and observed counts asymptotically follow the χ^2 distribution [144] with $k-1$ degrees of freedom, which allows computing a p-value for testing the null hypothesis [143].

Despite their usefulness, the KM and log-rank methods have a number of potential pitfalls and limitations. First, the interpretation of results is affected by the extent of the time axis, i.e., the last time point at which survival analysis is conducted [145]. Although it is possible to extend the analysis to the longest follow-up time, it is often better to exclude the most extreme time points because they have more statistical uncertainty and may be based on a limited number of subjects. In studies with a low number of total subjects, KM analysis is infeasible because the variance of survival estimates is too large (see Equation 2). When high-throughput molecular experiments are combined with KM analysis, care must be taken to avoid false discoveries. For example, one study found that 90% of transcription profiles of random sets having a hundred genes or more are statistically significant predictors of breast cancer outcome, and many published

Subject	Time t	δ
A	1	0
B	3	1
C	3	1
D	4	0
E	6	1
F	7	1
G	9	0



Event time t	Events	Censorings	Risk set: $T \geq t$	$\hat{S}(t)$
0	0	1	7 (A,B,C,D,E,F,G)	1
3	2	1	6 (B,C,D,E,F,G)	$1 \times \frac{4}{6} = 0.667$
6	1	0	3 (E,F,G)	$0.667 \times \frac{2}{3} = 0.444$
7	1	1	2 (F,G)	$0.444 \times \frac{1}{2} = 0.222$

Figure 7: Illustration of the Kaplan-Meier method. *Top row, left*: Raw generated survival data of seven subjects. Four subjects have an event and three are censored. The time scale for this example is arbitrary; for real data, it is often days, weeks or years. *Top row, right*: Visualization of Kaplan-Meier survival estimates for the illustration data. *Bottom row*: Processed survival data and numerical Kaplan-Meier survival function estimates. Data points have been aggregated into four time points corresponding to events; in addition, the zero time point is included to account for one censoring before the first event. For each time point, the second column shows the number of events at that time point. The third column shows the number of censorings between that and the next time point, i.e., during the half-open interval $[t_{(j)}, t_{(j+1)})$. The fourth column shows the subjects surviving until at least the given time. Finally, the last column shows the survival estimates computed using the rightmost part of Equation 1, noting that $\mathrm{P}(T > t_{(j)} | T \geq t_{(j)}) = \mathrm{P}(T > t_{(j)}) / \mathrm{P}(T \geq t_{(j)})$.

predictor gene sets are no better at predicting outcome than random [146]. Finally, careful interpretation of log-rank test results is needed when two survival curves are similar in some time points but different in others [145]. Log-rank computes the overall survival similarity by averaging differences over all time points, in this case in two qualitatively different time segments.

7 Results

In this thesis, I present six main results, which include three contributions to bioinformatics method development and three contributions to cancer biology (Table 5).

7.1 Anduril workflow framework

Anduril (`http://anduril.org`) is a software framework for implementing and executing complex analysis workflows. It is capable of executing completely automated bioinformatics workflows from sample importing and preprocessing, to statistical analysis, annotation with biodatabases, and generation of high-quality reports. Anduril provides scalability in both complexity dimensions: data volume and structural complexity. Anduril is extensible by third parties and can thus be adapted to rapidly changing technologies, such as novel deep sequencing analysis methods. All analyses in this thesis are executed with Anduril, and the methods developed in other publications have been integrated to Anduril.

Complexity dimensions: volume and structure

As a bioinformatics framework that provides multiple levels of scalability and extensibility, Anduril is intended to be used by researchers who are familiar with programming, such as R scripting. However, the reports generated by Anduril are targeted at bench biologists with no computational background. Such reports include colored Excel sheets for numeric data, generated web sites with searching and sorting functionality, and data visualization using multiple types of plots (Figure 8). Although the focus here is on biomedical applications, Anduril is a general-purpose framework that could be applied to data from other fields, such as "big data" generated by companies [147].

Complexity management: abstraction

Table 5: List of contributions in the thesis.

Publication	Type	Summary
Publication I	Methodological	Workflow framework for analyzing complex high-throughput data.
Publication I	Biomedical	Integrative analysis of transcriptomics, genomics, epigenomics and survival in GBM.
Publication II	Methodological	Algorithm for analyzing multi-marker time-series ChIP-seq data.
Publication II	Biomedical	Identification and elucidation of estrogen-responsive genes in breast cancer.
Publication III	Methodological	Software for flexible preprocessing of deep sequencing data; mathematical framework for encoding hypotheses.
Related Publ. I	Biomedical	Genome-wide analysis of AR and FOXA1 binding in prostate cancer, with clinical associations.

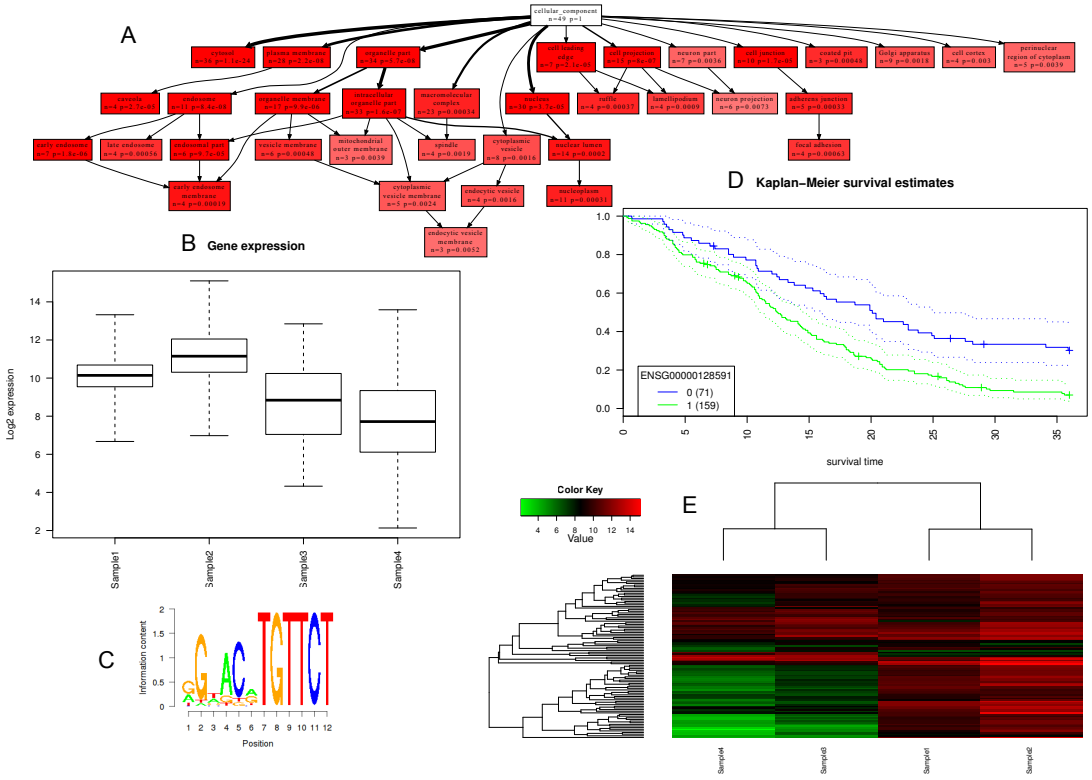


Figure 8: Examples of Anduril-generated reports. (A) Visualization of Gene Ontology [148] enrichment [149] results for members of the epidermal growth factor receptor pathway [150]; data from WikiPathways [151]. (B) Box plot of quantified gene expression values. (C) DNA consensus sequence logo [152]. (D) Kaplan-Meier survival curve; data from Publication I. (E) Heat map of quantified gene expression values.

Architecturally, Anduril is composed of three parts. First, Anduril *components* are executable software routines that each implement a well-defined part of an analysis, such as statistical inference. These components are important building blocks that can be reused between projects; thus, carefully designed and tested components reduce analysis time requirements. Anduril components implement a well-defined interface based on ports: a component reads data from one or more input ports (files) and writes results to one or more output ports. In addition to ports with complex data, simple parameters, such as integers, allow modulating component functionality. For the analyst, components are abstractions that hide implementation details and are accessed through an interface. Anduril components can be written in any programming language capable of reading and writing files; thus, Anduril allows reusing libraries in multiple languages, including R/Bioconductor [153], Java, Python and Matlab.

Complexity management: modularity

Second, *workflow construction* is the definition and configuration of a custom workflow for a particular project. This is done using a custom high-level domain-specific language, AndurilScript, designed for rapid construction of complex workflows. When executed, an AndurilScript program yields a workflow struc-

ture that is later validated and executed. The use of a domain-specific language is the most distinctive feature of Anduril, differing from graphical workflow construction (e.g., GenePattern and Taverna) and the use of general-purpose programming languages (e.g., Biopipe). The advantages of a custom language is integration of language features with workflow structure; the disadvantages are increased maintenance load for Anduril developers and a steeper learning curve for users. Nevertheless, AndurilScript is syntactically and semantically close to mainstream programming languages and thus relatively easy to learn. AndurilScript provides scalability for complex analyses through modularization: large workflows can be divided into parametrized sub-workflows. Common programming constructs, such as `if-else` statements and `for` loops, also increase workflow flexibility.

Complexity management: parallelism

The third architectural feature is *workflow execution* using a custom workflow engine. The Anduril engine supports parallel execution both within a single machine (threads) and distributed over a cluster (remote execution). Parallelism is limited by a user-configurable limit for processes, as well as natural constraints in the workflow topology. The engine is optimized for the iterative, "agile" nature of data analysis: results from all workflow steps are cached on disk, and upon a second execution of the workflow, only those steps are executed whose configuration has changed. This allows executing time-consuming preprocessing steps once and iteratively develop the downstream steps of statistics and report generation.

Computational overhead of Anduril workflow execution is relatively low, as Anduril is mainly used to launch external processes. Compared to manually executing analysis applications, Anduril overhead comes from managing the output files of each component and evaluating AndurilScript programs. The disk space occupied by components outputs can be significant; therefore, Anduril includes an option to automatically clean up output files as soon as they are not required by any further component.

7.2 Integrative analysis of heterogeneous "omics" data in GBM

Complexity dimensions: volume and structure

Understanding the complex molecular signature of glioblastoma is facilitated by genome-wide experiments that measure several molecular markers in parallel. Such experiments produce large-scale heterogeneous data that provide challenges for managing the bioinformatics analysis workflow as well as prioritizing statistical findings to find the "needle in the haystack".

In Publication I, we analyzed heterogeneous microarray data from The Cancer Genome Atlas (TCGA) [53] from 338 primary GBM patients to identify clinically relevant molecular markers. The microarrays included all publicly available TCGA data at the time of publication, including multiple types of expression arrays (exon, mRNA and miRNA), SNP arrays, array CGH and DNA methylation arrays. In addition, we used survival data for assessing clinical relevance.

Our answer to the challenge of managing the bioinformatics analysis was the use of Anduril; this study was, in addition to providing cancer research value, also intended to validate the applicability of Anduril to complex analysis projects. For finding the clinically relevant needle in the haystack of microarray data, we used a three-step strategy. First, we conducted statistical analysis on individual microarray platforms and integrated the results into gene-level summaries. This allows, for instance, identifying amplified genes that are overexpressed and also have hypomethylated promoters. Using results from the first step, we next conducted Kaplan-Meier survival analysis to identify genes that are candidate clinical markers and may have a functional role in cancer. Finally, we validated selected clinical markers *in vitro* in three GBM cell lines to assess whether the associated genes have functional roles in the cell. The results from the first and second steps are presented as a browsable web site that includes metrics such as gene expression changes and survival p-values, as well as related plots.

Our strongest result from step three is moesin (*MSN*), a gene coding for an actin cytoskeleton associated protein [154]. We found this gene to be overexpressed at the mRNA level and its overexpression had a negative effect on patient survival. In validation, *MSN* knockdown by siRNA resulted in decreased proliferation and increased apoptosis. Together, these features suggest an oncogenic function for *MSN* in GBM. Additional support for this hypothesis was provided by Zhu *et al.*, who elaborated the molecular function of *MSN* and further established the proliferative and clinical associations of the gene [155].

7.3 SPINLONG for complex ChIP-seq experiments

Complexity dimension: structure

SPINLONG (Spatial Pattern Identification by Non-Linear OptimizatioN with Global constraints; `http://csbi.ltdk.helsinki.fi/spinlong/`) is an algorithm for analyzing complex ChIP-seq and global run-on sequencing [107] (GRO-seq) data. SPINLONG is designed for ChIP targets that produce wide signal profiles, such as PolII and histone alterations, and is thus complementary to approaches that are optimized for sharp TF peaks [156].

SPINLONG has an expressive configuration schema that supports *in vitro* time series experiments and simultaneous analysis of multiple ChIP targets. As a configurable algorithm, SPINLONG has several use cases. A basic use case of SPINLONG is gene classification into transcribed, induced or repressed genes based on the state of the epigenome and transcription machinery. SPINLONG also allows differential transcription analysis based on quantification of PolII occupancy in gene bodies. A more complex use case is using the detailed SPINLONG output metrics to estimate elongation speed of the PolII machinery [157]. A drawback of the broad applicability of SPINLONG is its conceptual complexity and laborious configuration.

SPINLONG uses a spatio-temporal analog in which successive base pairs in a genome form the spatial dimension and samples from different time points, if any, form the temporal dimension. Short reads are assigned to fixed-width

genomic bins, and these read counts are the input to the algorithm. Whereas in algorithms such as MACS [116] the spatial pattern to be searched – peak – is predefined, in SPINLONG it is configurable by the user. Multiple patterns can be searched simultaneously, and gene classification can be done based on the best match. Each pattern encodes a hypothesis, such as "a gene is activated at 10 minutes after stimulus", in machine-readable form. In SPINLONG, a sample denotes a distinct sequencing library that may be a time point in a time series of the same ChIP target; a biological replicate; or a sample from a different ChIP target.

A concrete example of a spatio-temporal pattern is the activation of PolII machinery as a function of time in a gene induced by a stimulus. At the pre-stimulus sample, there is little PolII binding along the inactive gene. Shortly after the stimulus, PolII starts accumulating at the promoter and a short distance towards the gene body. After a while, PolII complexes have progressed halfway through the gene and thus the first half of the gene body is occupied by PolII. Finally, transcription of the first transcripts has been completed, and the gene body is fully occupied by PolII.

In SPINLONG, patterns are defined in the context of a pre-defined genomic region, such as the body of a gene; multiple independent regions are processed in parallel. Each region is divided by the user into one or more *segments*, which are contiguous genomic intervals with a homogeneous short read distribution. Specifically, each segment is expected to contain either a "high" or "low" amount of reads, as specified by the user. SPINLONG dynamically selects thresholds for low and high counts. Each sample has an independent division into segments. For increased expressiveness, there can be linear constraints between the lengths of segments, either within a single sample or between different samples.

Complexity management: parallelism

Based on the segment configuration and constraints defined by the user, the SPINLONG runtime optimization algorithm assigns lengths to each segment so that actual read counts within segment boundaries most accurately match the expected (low or high) counts. The segment length vector producing the optimal score is the "raw" result of the method. Pattern scoring is configurable by the user: for example, certain time points or ChIP targets can be given higher weight. The optimization algorithm is parallelized to speed up processing large data sets.

Scores and segment lengths are used in subsequent analysis steps to obtain answers to the higher-level research question. Gene classification is done by matching multiple patterns simultaneously, such as for induced and repressed (deactivated) genes, and assigning the class based on the highest score. Transcription quantification is done by counting reads in the segment that corresponds to the actively transcribed gene body; this allows adjusting the region to account for alternative promoters. Likewise, PolII elongation speed can be estimated from segment lengths corresponding to "high" count of PolII in successive time points.

7.4 Estrogen early response genes in breast cancer

Over half of breast cancers over-express the receptor for estrogen, ER(α), and are dependent on hormone signaling for proliferation [44]. Nevertheless, the cellular effects of estrogen are poorly understood. ER is a transcription factor having both genomic and nongenomic pathways for modulating transcription and cellular phenotype [45]. Thus, there are several methods for measuring the effects of estrogen on breast cancer cells. An often used *in vitro* method is mRNA profiling, in which the RNA products are measured after a delay of several hours to a few days after estrogen stimulation. The challenge with this approach is the selection of the appropriate time point(s): short genes are transcribed faster than long genes, and secondary transcriptional responses may be present by the time long genes have finished the primary transcription. A more direct measure of transcription modulation is the activity of PolII and certain histone alterations, which occur rapidly after the stimulus. PolII binding can be measured using ChIP-seq; alternatively, the nascent RNA strands can be profiled using GRO-seq.

Complexity dimension: structure

Using ChIP-seq, we measured PolII, H3K4me3, H2A.Z and ER activity in the MCF-7 breast cancer cell line [122] at time points 0, 10, 20, 40, 80, 160, 320, 640 and 1280 minutes after estradiol stimulus. Our goal was to identify early response genes and to assess their clinical relevance in an independent TCGA breast cancer cohort [54]. Analyzing a time-series ChIP-seq experiment measuring several markers in parallel is methodologically challenging, as most existing software are tuned for single samples of TF-like ChIP profiles. To address this challenge, we used the SPINLONG algorithm, which is scalable to complex ChIP-seq experiment designs.

We identified 777 estradiol early response genes, of which 699 are induced and 78 are repressed. Interestingly, many of these genes show a response in PolII binding patterns already at 10 minutes after stimulus, which implies rapid signal transduction using either the genomic or nongenomic pathway, and subsequent modification of chromatin amendable for transcription. Possibly, some of the early response genes are in a readiness state, with pre-assembled PolII complexes at the promoter. When comparing our 777 genes against published estrogen response gene sets from literature [158, 159, 160], we observed low overlaps both between our set and sets from literature, and between sets from different publications. One reason for the low overlaps is experimental design, as most published gene sets are based on mRNA profiling at a relatively late time point (e.g., 24 hours). This highlights the difficulty of unambiguously describing the "estrogen response": which experimental method should be selected to provide a canonical response gene set?

To filter the haystack of 777 genes, we searched for survival associations in the gene expression data from TCGA. The strongest result from this analysis was *ATPase Family, AAA Domain Containing 3B* (*ATAD3B*), which is induced in MCF-7 cells and whose overexpression is associated with decreased survival.

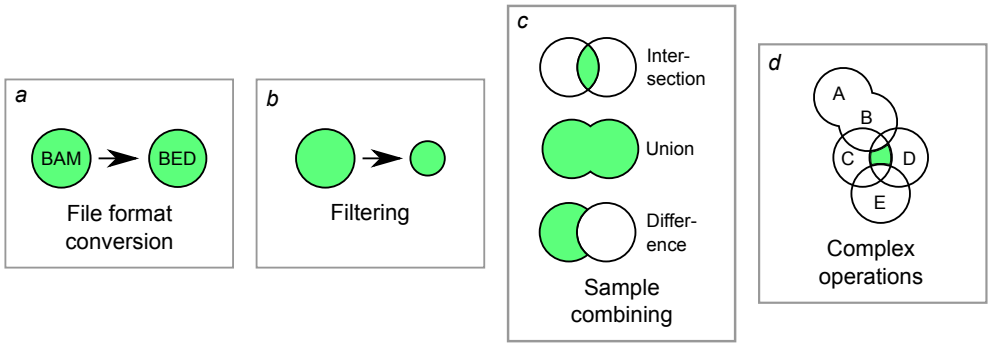


Figure 9: Use cases for GROK. Circles represent sets of genomic regions, which can be derived from aligned sequencing reads or output of sequence analysis algorithms. (a) File format conversion, in which core content of data remains unchanged. In this example, the conversion is from BAM to BED. (b) Filtering of genomic data, such as quality control for sequencing reads. (c) Combining genomic regions from several samples using elementary set operations. (d) Combining samples using a complex formula. In this example, the result set (green) is the intersection of samples C and D from which E and the union of A and B are removed, i.e., $((\mathrm{C} \cap \mathrm{D}) \setminus \mathrm{E}) \setminus (\mathrm{A} \cup \mathrm{B})$.

Induction of transcription led to increased mRNA levels as profiled using RNA-seq. We validated the survival association in two further published cohorts [123, 124]. *ATAD3B* is a mitochondrial protein whose molecular function in not known in detail; its role in tumorigenesis has previously been suggested [161].

7.5 GROK for flexible processing of deep sequencing data

Bioinformatics methods for analyzing deep sequencing data are developing rapidly and the tools are rarely standardized into formal workflows. Rather, state-of-the-art tools are often accessed from the command line or programming frameworks, such as Bioconductor, and combined into usable workflows using *ad hoc* scripting. This is aided by a degree of standardization in sequencing file formats, such as BAM [162], FASTQ [163] and VCF [164]. In these *ad hoc* workflows, there is often a need for "glue tools" that enable binding successive workflow step together. Such tools provide functionality for file format conversion, short read filtering and transformation, and sample comparison and combination (Figure 9). Published glue tools for scripting purposes include the BEDTools family [165, 166], BEDOPS [167] and CGAT [168].

Complexity management: abstraction

We took a systematic approach to constructing a flexible and extensible deep sequencing glue tool. First, we developed a mathematic formalism based on set algebra [169] that allows succinctly expressing analysis goals such as "select ChIP-seq peaks present in both samples A and B that do not overlap with peaks in sample C". This is an example of using abstraction for managing complexity in experimental design: a suitable formal language simplifies analysis specification and communication by providing high-level constructs for experimental features.

Based on the formalism, we designed and developed a practical sequencing toolkit, Genomic Region Operation Kit (GROK; http://csbi.ltdk.helsinki.fi/grok/), that implements the mathematical operations and supports various use cases for glue tools. Compared to other tools, GROK has a more orthogonal design: a relatively rich set of elementary operations are efficiently implemented, and complex operations are built by combining these. In addition, GROK is not bound to a specific programming interface, as it supports multiple programming languages (R, Python and Lua) as well as the command line.

7.6 AR and FOXA1 in prostate cancer

Analogously to estrogen in breast cancer, the majority of prostate tumors are dependent on androgen signaling for proliferation [43]. Androgens establish their transcriptional modulation mainly through the transcription factor AR, whose downstream targets, however, are not known in detail. We followed a similar general strategy as in the breast cancer study by first identifying molecular markers genome-wide *in vitro* and then assessing their clinical relevance using patient samples.

Using modified LNCaP human prostate cancer cells [125], we first measured AR binding sites (ARBSs) two hours after stimulation with the most potent natural androgen, dihydrotestosterone (DHT), using ChIP-seq. By analyzing the >8000 ARBSs identified using MACS [116], we identified enrichment of FOXA1 binding motifs within ARBS sequences. FOXA1 is a TF in the forkhead family that is associated to growth and differentiation of various tissues, including prostate and breast [170]. It is a co-regulator of expression together with other TFs, including AR [171]. In particular, FOXA1 is known to be a pioneer factor that binds to chromatin before other TFs. Thus, we focused on the AR–FOXA1 interaction by silencing FOXA1 expression by siRNA and conducting ChIP-seq for AR and FOXA1 in both parental (non-silenced) and FOXA1 silenced cells.

We found that FOXA1 has three distinct roles in co-regulating transcription with AR. Some ARBSs are independent from FOXA1, i.e., they are present in both parental and silenced cells. Some require the presence of FOXA1: they disappear in silenced cells, which corresponds to the pioneering functionality of FOXA1. Finally, a substantial number of ARBSs are masked by FOXA1 and only appear in silenced cells. The findings were validated with ChIP followed by quantitative polymerase chain reaction (ChIP-qPCR), and in a another prostate cancer cell line, VCaP. We further measured gene expression in both cell types and observed three transcriptional profiles that are consistent with the three classes of FOXA1–AR interaction. These results indicate that FOXA1 has a more complex interaction with AR than previously thought, as in addition to the pioneer role, it has two additional roles that are also reflected at the RNA level.

To assess the clinical relevance of FOXA1, we constructed a cohort of 350 prostate cancer patients to find correlations between protein levels and survival.

As a positive control, we verified that high AR protein expression is associated with reduced survival. Our main clinical finding was that high FOXA1 protein levels are associated with reduced survival, and low FOXA1 levels increase survival even in the presence of high AR protein levels. Together, these results indicate that FOXA1 has a multifaceted and clinically relevant role in prostate cancer.

8 Discussion

Malignant cells have numerous alterations in their genomes and epigenomes that alter the cellular phenotype in a heterogeneous and complex fashion. Identifying these alterations and their causal role in tumorigenesis forms a basis for more accurate diagnosis and prognosis, and improved therapeutic regimens, such as rational development of targeted anti-cancer drugs. As cancer cells have undergone mutations in various genomic regions, genome-wide measurement techniques are required for obtaining a holistic view of tumorigenic processes.

We developed software at multiple conceptual levels for scalable analysis of structurally complex genome-wide experimental data. The Anduril framework (Publication I) forms an umbrella system that manages the overall structure of analysis programs. It obtains scalability by parallelizing workflow execution, by leveraging libraries in multiple programming ecosystems in an integrated fashion, and by using a domain-specific language for constructing complex workflows. Anduril has enabled the analysis of several large-scale biomedical data sets. For instance, in Publication I, using a team of analysts, we constructed a large GBM analysis workflow that consisted of 350 distinct steps, and we have since extended the approach to three other TCGA-supplied cancers (`http://anduril.org`). So far, Publication I has been cited 43 times (Thomson Reuters, Web of Science), including in two commentaries [172, 173].

Complexity management: parallelism

Complexity management: modularity

The other methodological contributions, SPINLONG (Publication II) and GROK (Publication III), provide new workflow building blocks for analyzing complex ChIP-seq data and flexible preprocessing of deep sequencing data, respectively. SPINLONG expands the scope of ChIP-seq analysis algorithms by supporting multiple ChIP targets and time series experiments. GROK is based on a mathematical language that allows formulating common preprocessing tasks in a succinct manner.

Complexity dimension: structure

Complexity management: abstraction

Our three experimental contributions (Publication I, Publication II and Related Publication I) highlight three useful strategies in high-throughput experiments. In all experiments, we used both patient and cell line material, as these distinct approaches support each other. In patient material, we integrated molecular and survival data to combine the microscopic (cellular) level with the macroscopic (individual). In the GBM study, we started with patient material and validated results in cell lines. In the breast and prostate cancer studies, we followed an opposite strategy by conducting primary experiments in cell lines and assessing clinical relevance in patient cohorts.

In all publications, we started with genome-wide measurements whose initial analysis provided large result sets of genes or genomic regions, and then focused on one or a few genes as the primary result. These focus genes are *MSN* (Publication I), *ATAD3B* (Publication II) and *FOXA1* (Related Publication I). Focusing on a few genes makes the results more comprehensible for humans by reducing complexity, although it may also hide interesting details of the other

genes. To compensate for this, we also released the full results in tabular and WWW formats, using the reporting facilities of Anduril.

In the GBM and breast cancer studies, we utilized publicly available data sets from TCGA [53, 54] and other literature sources in addition to custom experiments. Other public data sources relevant for cancer research include the Encyclopedia of DNA Elements (ENCODE) [67] and International Cancer Genome Consortium (ICGC) [174]. Such data sets allow economical access to large patient cohorts, which facilitates deriving robust statistical conclusions on patient survival and molecular features of tumors. They also often use standardized laboratory protocols to enable comparison of experiments from several laboratories. On the other hand, public data sets are subject to competitive analysis from the worldwide biomedical community, which emphasizes early access to data and rapid development of analysis workflows. For the latter, Anduril can speed up development due to the reuse of existing analysis tools and workflows, and a workflow construction environment designed for programmers.

Complexity dimension: volume

In this thesis, we have focused the experimental and development efforts at the genomic, epigenomic and transcriptomic levels. It is important to remember, however, that much of cellular phenotypes, including malignancy, are established at the proteomic level [175]. Phenomena such as post-translational modifications and protein degradation are not observable at the DNA or RNA levels. High-throughput measurement techniques for nucleotides, such as deep sequencing and DNA microarrays, are more mature technologically than the corresponding techniques for proteomics, such as mass spectrometry [176] and protein arrays [177]. This is because nucleotide polymers are more amendable than proteins for sequencing and quantifying due to base pairing. However, as high-throughput proteomics technologies mature, they are expected to provide a new data layer for genome-wide cancer experiments. Anduril can be extended to support integrative analysis of such new data layers. Indeed, Anduril has already been extended for analysis of fluorescence activated cell sorting (FACS) data [178] and microscopy images (Ville Rantanen *et al.*, in preparation, `http://anduril.org/anima`).

In the future, high-throughput measurement technologies, such as deep sequencing, continue to be refined, and new technologies, such as third generation sequencing [179], will be deployed. These will not replace earlier technologies: for example, low-throughput PCR continues to be useful for targeted experiments and validation. As technology matures, the data generation step (e.g., sequencing) becomes a commodity and the bottleneck is in sample preparation and, in particular, data analysis and interpretation. New bioinformatics methods are needed for analyzing complex data sets [180], integrating heterogeneous data, and combining results from custom experiments with the literature. The most challenging step in high-throughput data analysis is identifying the needle in the haystack of results, i.e., obtaining biomedically relevant conclusions.

As data volume and complexity continues to grow, both in custom and published data sets, methods based on artificial intelligence (AI) [181, 182] may be useful

for aiding researchers in interpreting results. Computers are able to process large amounts of data, but often their internal model of the data is rudimentary: for example, a gene may be considered as a non-structured "black box". In contrast, an AI based analysis system could have a more flexible internal data model, which would help to integrate experimental data and make inferences in ways that are not as rigidly constrained as for one-purpose analysis algorithms. An early example of this paradigm is the Watson knowledge base system (IBM, New York, USA), which aids medical doctors in diagnosing and treating patients [183]. In basic biomedical research, an AI system could integrate a literature knowledge base with analysis and planning algorithms to flexibly explore data and suggest novel testable hypotheses to researchers.

Acknowledgements

This work was carried out in the Systems Biology Laboratory at the Medical Faculty of the University of Helsinki during 2007-2014. I thank Prof. Sampsa Hautaniemi for leading the laboratory in a professional manner and providing the right balance of active guidance and individual freedom in research. Sampsa originally helped me to transition from computer science to bioinformatics and biomedical research, and has continuously supported my development. In addition to his scientific competence, Sampsa is easy to approach, and can handle both pleasant small talk and constructive discussion of challenges in projects.

The members of my thesis committee, Prof. Juho Rousu and Prof. Jukka Westermarck, gave me helpful feedback on the progress of the thesis, as well as tips for career development. Outside the committee, Jukka was also a co-author in two of my publications, having an integral role in providing biomedical expertise.

I thank the reviewers of this thesis, Prof. Matti Nykter and Prof. Jorma Palvimo, for helpful suggestions for improving the text.

Our lab is an excellent place to work, with a friendly atmosphere and competent people. I therefore thank all current and past members of the Hautaniemi lab: Alejandra, Ali, Amjad, Anna-Maria, Chengyu, Chiara, Elena, Erkka, Janne, Javier, Jimmy, Julia, Kari, Katherine, Lauri, Lilli, Maninder, Marko, Mikko, Miko, Minna, Ping, Rainer, Riku, Roman, Rony, Sirkku, Tiia, Tuomo, Ville, Vladimir and Zitong. In particular, Marko's enthusiasm was integral during the early days of Anduril development. I thank Tiia, Javier, Riku and Ping for reading the manuscript of the thesis.

As computational biology is a highly interdisciplinary field, I have enjoyed collaboration and social occasions with numerous people from varying backgrounds, without whom no results would be possible. Interacting with you has also enriched my work life. I thank: Prof. Olli Jänne and Biswajyoti Sahu; Prof. Annamari Ranki, Emilia Carlsson, Pilvi Maliniemi and Sonja Hahtola; Prof. Pirkko Vihko and César Araujo; Prof. Lauri Aaltonen, Mervi Aavikko, Iikki Donner, Eevi Kaasinen, Riku Katainen, Johanna Kondelin and Esa Pitkänen; Prof. Eero Pukkala and Miia Artama; Dr. Matjaz Barboric and Andrii Bugai. In addition, I enjoyed lab visits at Prof. Henk Stunnenberg's lab (Nijmegen, the Netherlands) and Dr. Paul Meltzer's lab (Bethesda, MD, USA).

For funding, I thank the Helsinki Biomedical Graduate Program, Biomedicum Helsinki Foundation and the Cancer Society of Finland. Computational infrastructure was provided by IT Center for Science (CSC) and Institute for Molecular Medicine Finland (FIMM). At CSC, I thank my main contact, Danny Sternkopf; I also had interesting discussions with members of the Chipster team, Dr. Eija Korpelainen and Aleksi Kallio.

Finally, I thank my family, Markus, Pirkko and Risto, for being there.

Kristian Ovaska
Helsinki, 10.3.2014

Conflict of interest

Kristian Ovaska is a shareholder and board member in Significo Research Ltd, which sells services implemented using the open source Anduril software.

References

Page(s)

[1] Johnson, N. (2009) *Simply complexity; A clear guide to complexity theory*. (Oneworld Publications). 1, 2, 3, 4

[2] Alberts, B, Johnson, A, Lewis, J, Raff, M, Roberts, K, & Walter, P. (2002) *Molecular biology of the cell*. (Garland Science), Fourth edition. 1, 2, 4, 12, 13, 14

[3] Hanahan, D & Weinberg, R. A. (2011) Hallmarks of cancer: the next generation. *Cell* **144**, 646–674. 1, 2, 7, 8, 9, 10, 11, 12

[4] Weinberg, R. A. (2013) *The biology of cancer*. (Garland Science New York), Second edition. 1, 6, 8, 9, 10, 11, 12

[5] World Health Organization. (2013) Fact sheet 297: Cancer. 1, 6, 7

[6] Saiki, R. K, Gelfand, D. H, Stoffel, S, Scharf, S. J, Higuchi, R, Horn, G. T, Mullis, K. B, & Erlich, H. A. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487–491. 1

[7] Towbin, H, Staehelin, T, & Gordon, J. (1979) Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proceedings of the National Academy of Sciences of the USA* **76**, 4350–4354. 1

[8] Metzker, M. (2009) Sequencing technologies–the next generation. *Nature Reviews Genetics* **11**, 31–46. 1, 19, 20

[9] Barrett, J. C & Kawasaki, E. S. (2003) Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression. *Drug Discovery Today* **8**, 134–141. 1, 16, 17

[10] Diener, E, Emmons, R. A, Larsen, R. J, & Griffin, S. (1985) The satisfaction with life scale. *Journal of Personality Assessment* **49**, 71–75. 1

[11] Bird, I. (2011) Computing for the Large Hadron Collider. *Annual Review of Nuclear and Particle Science* **61**, 99–118. 1

[12] Tyson, J. J, Baumann, W. T, Chen, C, Verdugo, A, Tavassoly, I, Wang, Y, Weiner, L. M, & Clarke, R. (2011) Dynamic modelling of oestrogen signalling and cell fate in breast cancer cells. *Nature Reviews Cancer* **11**, 523–532. 2

[13] Harrow, J, Frankish, A, Gonzalez, J. M, Tapanari, E, Diekhans, M, Kokocinski, F, Aken, B. L, Barrell, D, Zadissa, A, Searle, S, et al. (2012) GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research* **22**, 1760–1774. 2

[14] Wu, J, Vallenius, T, Ovaska, K, Westermarck, J, Mäkelä, T. P, & Hautaniemi, S. (2009) Integrated network analysis platform for protein-protein interactions. *Nature Methods* **6**, 75–77. 2

[15] Doulatov, S, Notta, F, Laurenti, E, & Dick, J. E. (2012) Hematopoiesis: a human perspective. *Cell Stem Cell* **10**, 120–136. 2

[16] Grizzi, F & Chiriva-Internati, M. (2006) Cancer: looking for simplicity and finding complexity. *Cancer Cell International* **6**, 4. 2, 6

[17] Malkin, D, Li, F. P, Strong, L. C, Fraumeni, J, Nelson, C. E, Kim, D. H, Kassel, J, Gryka, M. A, Bischoff, F. Z, Tainsky, M. A, et al. (1990) Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* **250**, 1233–1238. 2

[18] Levine, A. J & Oren, M. (2009) The first 30 years of p53: growing ever more complex. *Nature Reviews Cancer* **9**, 749–758. 2, 10

[19] Tatusov, R. L, Fedorova, N. D, Jackson, J. D, Jacobs, A. R, Kiryutin, B, Koonin, E. V, Krylov, D. M, Mazumder, R, Mekhedov, S. L, Nikolskaya, A. N, et al. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41. 2

[20] Gottesman, M. M. (2002) Mechanisms of cancer drug resistance. *Annual Review of Medicine* **53**, 615–627. 3

[21] Clarke, R, Liu, M. C, Bouker, K. B, Gu, Z, Lee, R. Y, Zhu, Y, Skaar, T. C, Gomez, B, O'Brien, K, Wang, Y, et al. (2003) Antiestrogen resistance in breast cancer and the role of estrogen receptor signaling. *Oncogene* **22**, 7316–7339. 3, 7, 9

[22] Hazen, R. M, Griffin, P. L, Carothers, J. M, & Szostak, J. W. (2007) Functional information and the emergence of biocomplexity. *Proceedings of the National Academy of Sciences of the USA* **104**, 8574–8581. 3

[23] Gell-Mann, M. (1995) What is complexity? Remarks on simplicity and complexity by the Nobel Prize-winning author of The Quark and the Jaguar. *Complexity* **1**, 16–19. 3

[24] Cormen, T, Leiserson, C. E, Rivest, R. L, & Stein, C. (2001) *Introduction to algorithms*. (The MIT Press), Second edition. 3

[25] Tucker, A. B. (2004) *Computer Science handbook*. (CRC Press), Second edition. 3

[26] Pržulj, N, Corneil, D. G, & Jurisica, I. (2004) Modeling interactome: scale-free or geometric? *Bioinformatics* **20**, 3508–3515. 3

[27] Pan, Q, Shai, O, Lee, L. J, Frey, B. J, & Blencowe, B. J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* **40**, 1413–1415. 3

[28] Kolmogorov, A. (1965) Three approaches to the quantitative definition of information. *Problems of Information Transmission* **1**, 1–7. 4

[29] Pinho, A. J, Ferreira, P. J, Neves, A. J, & Bastos, C. A. (2011) On the representability of complete genomes by multiple competing finite-context (Markov) models. *PLoS ONE* **6**, e21588. 4

[30] Sutter, H. (2005) The free lunch is over: A fundamental turn toward concurrency in software. *Dr. Dobb's Journal* **30**, 202–210. 4

[31] Barabási, A.-L & Oltvai, Z. N. (2004) Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* **5**, 101–113. 4

[32] Sullivan, K. J, Griswold, W. G, Cai, Y, & Hallen, B. (2001) *The structure and value of modularity in software design*. (ACM), Vol. 26, pp. 99–108. 4

[33] Campbell, N. A, Reece, J. B, Urry, L. A, Cain, M. L, Wasserman, S. A, Minorsky, P. V, & Jackson, R. B. (2008) *Biology*. (Pearson Benjamin Cummings), Eighth edition. 4

[34] Jones, M. R. (2005) *Idealization and abstraction: A framework*. (Rodopi Amsterdam) Vol. 86, pp. 173–217. 4, 5

[35] Hanahan, D & Weinberg, R. A. (2000) The hallmarks of cancer. *Cell* **100**, 57–70. 6, 7

[36] Gerlinger, M, Rowan, A. J, Horswell, S, Larkin, J, Endesfelder, D, Gronroos, E, Martinez, P, Matthews, N, Stewart, A, Tarpey, P, et al. (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England Journal of Medicine* **366**, 883–892. 6

[37] Furnari, F. B, Fenton, T, Bachoo, R. M, Mukasa, A, Stommel, J. M, Stegh, A, Hahn, W. C, Ligon, K. L, Louis, D. N, Brennan, C, et al. (2007) Malignant astrocytic glioma: genetics, biology, and paths to treatment. *Genes & Development* **21**, 2683–2710. 6, 7, 10, 11, 12

[38] Johansson, J.-E, Holmberg, L, Johansson, S, Bergström, R, & Adami, H.-O. (1997) Fifteen-year survival in prostate cancer. *The Journal of the American Medical Association* **277**, 467–471. 6

[39] Ford, D, Easton, D, Stratton, M, Narod, S, Goldgar, D, Devilee, P, Bishop, D, Weber, B, Lenoir, G, Chang-Claude, J, et al. (1998) Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. *The American Journal of Human Genetics* **62**, 676–689. 6

[40] Tavassoli, F. A & Devilee, P. (2003) *Pathology and genetics of tumours of the breast and female genital organs*. (World Health Organization). 7

[41] Bauer, K. R, Brown, M, Cress, R. D, Parise, C. A, & Caggiano, V. (2007) Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype. *Cancer* **109**, 1721–1728. 7

[42] Grönberg, H. (2003) Prostate cancer epidemiology. *The Lancet* **361**, 859–864. 7

[43] Friedlander, T. W & Ryan, C. J. (2012) Targeting the Androgen Receptor. *Urologic Clinics of North America* **39**, 453–464. 7, 8, 9, 37

[44] Ali, S & Coombes, R. C. (2000) Estrogen receptor alpha in human breast cancer: occurrence and significance. *Journal of Mammary Gland Biology and Neoplasia* **5**, 271–281. 8, 35

[45] Levin, E. R. (2001) Invited Review: Cell localization, physiology, and nongenomic actions of estrogen receptors. *Journal of Applied Physiology* **91**, 1860–1867. 8, 9, 14, 35

[46] Matsumoto, T, Sakari, M, Okada, M, Yokoyama, A, Takahashi, S, Kouzmenko, A, & Kato, S. (2013) The androgen receptor in health and disease. *Annual Review of Physiology* **75**, 201–224. 8

[47] Guerriero, G. (2009) Vertebrate sex steroid receptors: evolution, ligands, and neurodistribution. *Annals of the New York Academy of Sciences* **1163**, 154–168. 9

[48] den Hollander, P, Savage, M. I, & Brown, P. H. (2013) Targeted therapy for breast cancer prevention. *Frontiers in Oncology* **3**, 250. 9

[49] Risbridger, G. P, Davis, I. D, Birrell, S. N, & Tilley, W. D. (2010) Breast and prostate cancer: more similar than different. *Nature Reviews Cancer* **10**, 205–212. 9

[50] Bianconi, E, Piovesan, A, Facchin, F, Beraudi, A, Casadei, R, Frabetti, F, Vitale, L, Pelleri, M. C, Tassani, S, Piva, F, et al. (2013) An estimation of the number of cells in the human body. *Annals of Human Biology* pp. 1–9. 9

[51] Harbour, J. W & Dean, D. C. (2000) The Rb/E2F pathway: expanding roles and emerging paradigms. *Genes & Development* **14**, 2393–2409. 10

[52] Kandoth, C, McLellan, M. D, Vandin, F, Ye, K, Niu, B, Lu, C, Xie, M, Zhang, Q, McMichael, J. F, Wyczalkowski, M. A, et al. (2013) Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339. 10, 13

[53] McLendon, R, Friedman, A, Bigner, D, Van Meir, E, Brat, D, Mastrogianakis, G, Olson, J, Mikkelsen, T, Lehman, N, Aldape, K, et al. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068. 10, 25, 32, 40

[54] The Cancer Genome Atlas Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70. 10, 22, 25, 35, 40

[55] Chi, S.-G, deVere White, R. W, Meyers, F. J, Siders, D. B, Lee, F, & Gumerlock, P. H. (1994) p53 in prostate cancer: frequent expressed transition mutations. *Journal of the National Cancer Institute* **86**, 926–933. 10

[56] Muller, P. A & Vousden, K. H. (2013) p53 mutations in cancer. *Nature Cell Biology* **15**, 2–8. 10

[57] Moyzis, R, Buckingham, J, Cram, L, Dani, M, Deaven, L, Jones, M, Meyne, J, Ratliff, R, & Wu, J. (1988) A highly conserved repetitive DNA sequence, $(TTAGGG)_n$, present at the telomeres of human chromosomes. *Proceedings of the National Academy of Sciences of the USA* **85**, 6622–6626. 11

[58] Poole, J. C, Andrews, L. G, & Tollefsbol, T. O. (2001) Activity, function, and gene regulation of the catalytic subunit of telomerase (hTERT). *Gene* **269**, 1–12. 11

[59] Nabetani, A & Ishikawa, F. (2011) Alternative lengthening of telomeres pathway: recombination-mediated telomere maintenance mechanism in human cells. *Journal of Biochemistry* **149**, 5–14. 11

[60] Shay, J & Bacchetti, S. (1997) A survey of telomerase activity in human cancer. *European Journal of Cancer* **33**, 787–791. 11

[61] Baeriswyl, V & Christofori, G. (2009) The angiogenic switch in carcinogenesis. *Seminars in Cancer Biology* **19**, 329–337. 11

[62] Talmadge, J. E & Fidler, I. J. (2010) AACR centennial series: the biology of cancer metastasis: historical perspective. *Cancer Research* **70**, 5649–5669. 12

[63] Ciccia, A & Elledge, S. J. (2010) The DNA damage response: making it safe to play with knives. *Molecular Cell* **40**, 179–204. 12

[64] Drake, J. W, Charlesworth, B, Charlesworth, D, & Crow, J. F. (1998) Rates of spontaneous mutation. *Genetics* **148**, 1667–1686. 13

[65] Albertson, D. G, Collins, C, McCormick, F, & Gray, J. W. (2003) Chromosome aberrations in solid tumors. *Nature Genetics* **34**. 13

[66] Hager, G, McNally, J, & Misteli, T. (2009) Transcription dynamics. *Molecular Cell* **35**, 741–753. 13

[67] The ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74. 13, 22, 40

[68] Chen, W & Moore, M. J. (2014) The spliceosome: disorder and dynamics defined. *Current Opinion in Structural Biology* **24**, 141–149. 13

[69] Bode, A. M & Dong, Z. (2004) Post-translational modification of p53 in tumorigenesis. *Nature Reviews Cancer* **4**, 793–805. 14

[70] Wang, E. T, Sandberg, R, Luo, S, Khrebtukova, I, Zhang, L, Mayr, C, Kingsmore, S. F, Schroth, G. P, & Burge, C. B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476. 14

[71] Goldberg, A. D, Allis, C. D, & Bernstein, E. (2007) Epigenetics: a landscape takes shape. *Cell* **128**, 635–638. 14

[72] Laird, P. W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics* **11**, 191–203. 14

[73] Jones, P. A. (1999) The DNA methylation paradox. *Trends in Genetics* **15**, 34–37. 14

[74] Shukla, S, Kavak, E, Gregory, M, Imashimizu, M, Shutinoski, B, Kashlev, M, Oberdoerffer, P, Sandberg, R, & Oberdoerffer, S. (2011) CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**, 74–79. 14

[75] Berger, S. L. (2007) The complex language of chromatin regulation during transcription. *Nature* **447**, 407–412. 14, 15

[76] Jin, C, Zang, C, Wei, G, Cui, K, Peng, W, Zhao, K, & Felsenfeld, G. (2009) H3 3/H2A.Z double variant–containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nature Genetics* **41**, 941–945. 15

[77] Kawasaki, E. (2006) The end of the microarray tower of Babel: Will universal standards lead the way? *Journal of Biomolecular Techniques* **17**, 200. 16, 17

[78] Allison, D, Cui, X, Page, G, & Sabripour, M. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics* **7**, 55–65. 16

[79] Schena, M, Shalon, D, Davis, R, & Brown, P. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470. 16

[80] Hoheisel, J. D. (2006) Microarray technology: beyond transcript profiling and genotype analysis. *Nature Reviews Genetics* **7**, 200–210. 16, 17

[81] Malone, J. H & Oliver, B. (2011) Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology* **9**, 34. 16

[82] Park, P. J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* **10**, 669–680. 16, 21, 22, 23

[83] Albert, T. J, Molla, M. N, Muzny, D. M, Nazareth, L, Wheeler, D, Song, X, Richmond, T. A, Middle, C. M, Rodesch, M. J, Packard, C. J, et al. (2007) Direct selection of human genomic loci by microarray hybridization. *Nature Methods* **4**, 903–905. 16, 18

[84] Mujumdar, R, Ernst, L, Mujumdar, S, Lewis, C, & Waggoner, A. (1993) Cyanine dye labeling reagents: Sulfoindocyanine succinimidyl esters. *Bioconjugate Chemistry* **4**, 105–111. 17

[85] Chen, P, Lepikhova, T, Hu, Y, Monni, O, & Hautaniemi, S. (2011) Comprehensive exon array data processing method for quantitative analysis of alternative spliced variants. *Nucleic Acids Research* **39**, e123–e123. 17

[86] Gertz, E. M, Sengupta, K, Difilippantonio, M, Ried, T, & Schäffer, A. (2009) Evaluating annotations of an Agilent expression chip suggests that many features cannot be interpreted. *BMC Genomics* **10**, 566. 18

[87] Hutchison, C. A. (2007) DNA sequencing: bench to bedside and beyond. *Nucleic Acids Research* **35**, 6227–6237. 18, 20

[88] Watson, J. D & Crick, F. H. (1953) Molecular structure of nucleic acids. *Nature* **171**, 737–738. 18

[89] Sanger, F, Nicklen, S, & Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the USA* **74**, 5463–5467. 18

[90] Shendure, J & Ji, H. (2008) Next-generation DNA sequencing. *Nature Biotechnology* **26**, 1135–1145. 18, 19

[91] Lander, E. S, Linton, L. M, Birren, B, Nusbaum, C, Zody, M. C, Baldwin, J, Devon, K, Dewar, K, Doyle, M, FitzHugh, W, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921. 19, 22

[92] Venter, J. C, Adams, M. D, Myers, E. W, Li, P. W, Mural, R. J, Sutton, G. G, Smith, H. O, Yandell, M, Evans, C. A, Holt, R. A, et al. (2001) The sequence of the human genome. *Science* **291**, 1304–1351. 19, 22

[93] Battelle Technology Partnership Practice. (2013) The Impact of Genomics on the U.S. Economy. *United for Medical Research*. 19

[94] Illumina, Inc. (2012) HiSeq(R) 1500/2500 Sequencing Systems: Specification sheet. 19

[95] Illumina, Inc. (2013) MiSeq(R) System: Specification sheet. 19

[96] Illumina, Inc. (2006) Press release Nov. 13, 2006: Illumina signs definitive agreement to acquire Solexa. 19

[97] Bentley, D. R, Balasubramanian, S, Swerdlow, H. P, Smith, G. P, Milton, J, Brown, C. G, Hall, K. P, Evers, D. J, Barnes, C. L, Bignell, H. R, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59. 19, 20, 21

[98] Ewing, B & Green, P. (1998) Base-calling of automated sequencer traces using Phred II: Error probabilities. *Genome Research* **8**, 186–194. 21

[99] Langmead, B & Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359. 21

[100] Li, H & Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760. 21

[101] Smith, T. F & Waterman, M. S. (1981) Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195–197. 21

[102] Altschul, S. F, Madden, T. L, Schäffer, A. A, Zhang, J, Zhang, Z, Miller, W, & Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402. 21

[103] The 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65. 22

[104] Meyerson, M, Gabriel, S, & Getz, G. (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics* **11**, 685–696. 22

[105] Ellis, M. J, Ding, L, Shen, D, Luo, J, Suman, V. J, Wallis, J. W, Van Tine, B. A, Hoog, J, Goiffon, R. J, Goldstein, T. C, et al. (2012) Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486**, 353–360. 22

[106] Wang, Z, Gerstein, M, & Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57–63. 22

[107] Core, L, Waterfall, J, & Lis, J. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848. 22, 33

[108] Hah, N, Danko, C, Core, L, Waterfall, J, Siepel, A, Lis, J, & Kraus, W. (2011) A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* **145**, 622–634. 22

[109] Krueger, F, Kreck, B, Franke, A, & Andrews, S. R. (2012) DNA methylome analysis using short bisulfite sequencing data. *Nature Methods* **9**, 145–151. 22

[110] de Wit, E & de Laat, W. (2012) A decade of 3C technologies: insights into nuclear organization. *Genes & Development* **26**, 11–24. 22

[111] Ingolia, N. T, Ghaemmaghami, S, Newman, J. R, & Weissman, J. S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223. 22

[112] Wooley, J. C, Godzik, A, & Friedberg, I. (2010) A primer on metagenomics. *PLoS Computational Biology* **6**, e1000667. 22

[113] Barski, A, Cuddapah, S, Cui, K, Roh, T.-Y, Schones, D. E, Wang, Z, Wei, G, Chepelev, I, & Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837. 21

[114] Orlando, V. (2000) Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends in Biochemical Sciences* **25**, 99–104. 21, 22

[115] Wilbanks, E & Facciotti, M. (2010) Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE* **5**, e11471. 23

[116] Zhang, Y, Liu, T, Meyer, C, Eeckhoute, J, Johnson, D, Bernstein, B, Nussbaum, C, Myers, R, Brown, M, Li, W, et al. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137. 23, 34, 37

[117] Bryne, J, Valen, E, Tang, M, Marstrand, T, Winther, O, Da Piedade, I, Krogh, A, Lenhard, B, & Sandelin, A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Research* **36**, D102–D106. 23

[118] Bailey, T. L, Boden, M, Buske, F. A, Frith, M, Grant, C. E, Clementi, L, Ren, J, Li, W. W, & Noble, W. S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research* **37**, W202–W208. 23

[119] Ouyang, Z, Zhou, Q, & Wong, W. H. (2009) ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences of the USA* **106**, 21521–21526. 23

[120] Sun, H, Wu, J, Wickramasinghe, P, Pal, S, Gupta, R, Bhattacharyya, A, Agosto-Perez, F, Showe, L, Huang, T, & Davuluri, R. (2011) Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq. *Nucleic Acids Research* **39**, 190. 23

[121] Welboren, W, Van Driel, M, Janssen-Megens, E, Van Heeringen, S, Sweep, F, Span, P, & Stunnenberg, H. (2009) ChIP-Seq of ERα and RNA polymerase II defines genes differentially responding to ligands. *The EMBO Journal* **28**, 1418–1428. 23

[122] Soule, H, Vazquez, J, Long, A, Albert, S, & Brennan, M. (1973) A human cell line from a pleural effusion derived

from a breast carcinoma. *Journal of the National Cancer Institute* **51**, 1409–1416. 25, 35

[123] Miller, L, Smeds, J, George, J, Vega, V, Vergara, L, Ploner, A, Pawitan, Y, Hall, P, Klaar, S, Liu, E, et al. (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the USA* **102**, 13550–13555. 25, 36

[124] Pawitan, Y, Bjöhle, J, Amler, L, Borg, A.-L, Egyhazi, S, Hall, P, Han, X, Holmberg, L, Huang, F, Klaar, S, et al. (2005) Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research* **7**, R953. 25, 36

[125] Cleutjens, C, Steketee, K, van Eekelen, C, Van der Korput, J, Brinkmann, A, & Trapman, J. (1997) Both androgen receptor and glucocorticoid receptor are able to induce prostate-specific antigen expression, but differ in their growth-stimulating properties of LNCaP cells. *Endocrinology* **138**, 5293–5300. 25, 37

[126] Georgakopoulos, D, Hornick, M, & Sheth, A. (1995) An overview of workflow management: From process modeling to workflow automation infrastructure. *Distributed and Parallel Databases* **3**, 119–153. 26

[127] Romano, P. (2008) Automation of in-silico data analysis processes through workflow management systems. *Briefings in Bioinformatics* **9**, 57–68. 26

[128] Russell, N, ter Hofstede, A. H, van der Aalst, W. M, & Mulyar, N. (2006) Workflow control-flow patterns: A revised view. *BPM Center Report BPM-06-22*. 26, 27

[129] Asanovic, K, Bodik, R, Demmel, J, Keaveny, T, Keutzer, K, Kubiatowicz, J, Morgan, N, Patterson, D, Sen, K, Wawrzynek, J, et al. (2009) A view of the parallel computing landscape. *Communications of the ACM* **52**, 56–67. 26

[130] Russell, N, ter Hofstede, A. H, van der Aalst, W. M, & Mulyar, N. (2004) Workflow data patterns. *QUT Technical report, FIT-TR-2004-01*. 26, 27

[131] Gamma, E, Helm, R, Johnson, R, & Vlissides, J. (1995) *Design patterns: elements of reusable object-oriented software*. (Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA), First edition. 26

[132] Giardine, B, Riemer, C, Hardison, R, Burhans, R, Elnitski, L, Shah, P, Zhang, Y, Blankenberg, D, Albert, I, Taylor, J, et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Research* **15**, 1451–1455. 26

[133] Oinn, T, Greenwood, M, Addis, M, Alpdemir, M, Ferris, J, Glover, K, Goble, C, Goderis, A, Hull, D, Marvin, D, et al. (2006) Taverna: Lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience* **18**, 1067–1100. 26

[134] Reich, M, Liefeld, T, Gould, J, Lerner, J, Tamayo, P, & Mesirov, J. (2006) GenePattern 2.0. *Nature Genetics* **38**, 500–501. 26

[135] Kallio, M. A, Tuimala, J, Hupponen, T, Klemelä, P, Gentile, M, Scheinin, I, Koski, M, Käki, J, & Korpelainen, E. (2011) Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics* **12**, 507. 26

[136] Altintas, I, Berkley, C, Jaeger, E, Jones, M, Ludascher, B, & Mock, S. (2004) Kepler: an extensible system for design and execution of scientific workflows. *Proceedings of the 16th International Conference on Scientific and Statistical Database Management* pp. 423–424. 26

[137] Jagla, B, Wiswedel, B, & Coppée, J.-Y. (2011) Extending KNIME for next-generation sequencing data analysis. *Bioinformatics* **27**, 2907–2909. 26

[138] Hoon, S, Ratnapu, K. K, Chia, J.-m, Kumarasamy, B, Juguang, X, Clamp, M, Stabenau, A, Potter, S, Clarke, L, & Stupka, E. (2003) Biopipe: a flexible framework for protocol-based bioinformatics analysis. *Genome Research* **13**, 1904–1915. 26

[139] Deelman, E, Singh, G, Su, M.-H, Blythe, J, Gil, Y, Kesselman, C, Mehta, G, Vahi, K, Berriman, G. B, Good, J, et al. (2005) Pegasus: A framework for mapping complex scientific workflows onto distributed systems. *Scientific Programming* **13**, 219–237. 26

[140] Goodstadt, L. (2010) Ruffus: a lightweight Python library for computational pipelines. *Bioinformatics* **26**, 2778–2779. 26

[141] Kleinbaum, D. G & Klein, M. (2005) *Survival analysis*. (Springer), Second edition. 27, 28

[142] Kaplan, E. L & Meier, P. (1958) Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481. 27, 28

[143] Bland, J. M & Altman, D. G. (2004) Statistics notes: the logrank test. *British Medical Journal* **328**, 1073. 28

[144] Pearson, K. (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical*

Magazine Series **50**, 157–175. 28

[145] Pocock, S. J, Clayton, T. C, & Altman, D. G. (2002) Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. *The Lancet* **359**, 1686–1689. 28, 29

[146] Venet, D, Dumont, J. E, & Detours, V. (2011) Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Computational Biology* **7**, e1002240. 29

[147] Manyika, J, Chui, M, Brown, B, Bughin, J, Dobbs, R, Roxburgh, C, & Hung Byers, A. (2011) *Big data: The next frontier for innovation, competition, and productivity*. (McKinsey Global Institute). 30

[148] Ashburner, M, Ball, C, Blake, J, Botstein, D, Butler, H, Cherry, J, Davis, A, Dolinski, K, Dwight, S, Eppig, J, Harris, M, Hill, D, Issel-Tarver, L, Kasarskis, A, Lewis, S, Matese, J, Richardson, J, Ringwald, M, Rubin, G, & Sherlock, G. (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29. 31

[149] Ovaska, K, Laakso, M, & Hautaniemi, S. (2008) Fast Gene Ontology based clustering for microarray experiments. *BioData Mining* **1**, 11. 31

[150] Oda, K, Matsuoka, Y, Funahashi, A, & Kitano, H. (2005) A comprehensive pathway map of epidermal growth factor receptor signaling. *Molecular Systems Biology* **1**, 2005.0010. 31

[151] Pico, A. R, Kelder, T, van Iersel, M. P, Hanspers, K, Conklin, B. R, & Evelo, C. (2008) WikiPathways: pathway editing for the people. *PLoS Biology* **6**, e184. 31

[152] Schneider, T. D & Stephens, R. M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research* **18**, 6097–6100. 31

[153] Gentleman, R. C, Carey, V. J, Bates, D. M, et al. (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80. 31

[154] Bretscher, A, Edwards, K, & Fehon, R. G. (2002) ERM proteins and merlin: integrators at the cell cortex. *Nature Reviews Molecular Cell Biology* **3**, 586–599. 33

[155] Zhu, X, Morales, F. C, Agarwal, N. K, Dogruluk, T, Gagea, M, & Georgescu, M.-M. (2013) Moesin is a glioma progression marker that induces proliferation and Wnt/β-catenin pathway activation via interaction with CD44. *Cancer Research* **73**, 1142–1155. 33

[156] Pepke, S, Wold, B, & Mortazavi, A. (2009) Computation for ChIP-seq and RNA-seq studies. *Nature Methods* **6**, S22–S32. 33

[157] Selth, L, Sigurdsson, S, & Svejstrup, J. (2010) Transcript elongation by RNA polymerase II. *Annual Review of Biochemistry* **79**, 271–293. 33

[158] Tang, S, Han, H, & Bajic, V. B. (2004) ERGDB: estrogen responsive genes database. *Nucleic Acids Research* **32**, D533–D536. 35

[159] Jin, V, Sun, H, Pohar, T, Liyanarachchi, S, Palaniswamy, S, Huang, T. H, & Davuluri, R. (2005) ERTargetDB: an integral information resource of transcription regulation of estrogen receptor target genes. *Journal of Molecular Endocrinology* **35**, 225–230. 35

[160] Cicatiello, L, Mutarelli, M, Grober, O, Paris, O, Ferraro, L, Ravo, M, Tarallo, R, Luo, S, Schroth, G. P, Seifert, M, et al. (2010) Estrogen receptor α controls a gene network in luminal-like breast cancer cells comprising multiple transcription factors and microRNAs. *The American Journal of Pathology* **176**, 2113–2130. 35

[161] Li, S & Rousseau, D. (2012) ATAD3, a vital membrane bound mitochondrial ATPase involved in tumor progression. *Journal of Bioenergetics and Biomembranes* **44**, 189–197. 36

[162] Li, H, Handsaker, B, Wysoker, A, Fennell, T, Ruan, J, Homer, N, Marth, G, Abecasis, G, Durbin, R, et al. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. 36

[163] Cock, P, Fields, C, Goto, N, Heuer, M, & Rice, P. (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* **38**, 1767–1771. 36

[164] Danecek, P, Auton, A, Abecasis, G, Albers, C, Banks, E, DePristo, M, Handsaker, R, Lunter, G, Marth, G, Sherry, S, et al. (2011) The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158. 36

[165] Quinlan, A & Hall, I. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841. 36

[166] Dale, R, Pedersen, B, & Quinlan, A. (2011) Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424. 36

[167] Neph, S, Kuehn, M, Reynolds, A, Haugen, E, Thurman, R, Johnson, A, Rynes, E, Maurano, M, Vierstra, J, Thomas, S, et al. (2012) BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920.

36

[168] Sims, D, Ilott, N. E, Sansom, S. N, Sudbery, I. M, Johnson, J. S, Fawcett, K. A, Berlanga-Taylor, A. J, Luna-Valero, S, Ponting, C. P, & Heger, A. (2014) CGAT: computational genomics analysis toolkit. *Bioinformatics*. 36

[169] Bronshtein, I. N, Semendyayev, K. A, Musiol, G, & Muehlig, H. (2007) *Handbook of mathematics*. (Springer), Fifth edition. 36

[170] Kaestner, K. H. (2010) The FoxA factors in organogenesis and differentiation. *Current Opinion in Genetics & Development* **20**, 527–532. 37

[171] Gao, N, Zhang, J, Rao, M. A, Case, T. C, Mirosevich, J, Wang, Y, Jin, R, Gupta, A, Rennie, P. S, & Matusik, R. J. (2003) The role of hepatocyte nuclear factor-3α (Forkhead Box A1) and androgen receptor in transcriptional regulation of prostatic genes. *Molecular Endocrinology* **17**, 1484–1507. 37

[172] Almeida, J. S et al. (2010) Computational ecosystems for data-driven medical genomics. *Genome Medicine* **2**, 67. 39

[173] Evans, C. (2011) Scientists develop new database that provides comprehensive view of glioblastoma multiforme genome. *TCGA Research Briefs*. 39

[174] Hudson, T. J, Anderson, W, Aretz, A, Barker, A. D, Bell, C, Bernabé, R. R, Bhan, M, Calvo, F, Eerola, I, Gerhard, D. S, et al. (2010) International network of cancer genome projects. *Nature* **464**, 993–998. 40

[175] Yaffe, M. B. (2013) The scientific drunk and the lamppost: massive sequencing efforts in cancer discovery and treatment. *Science Signaling* **6**, pe13. 40

[176] Aebersold, R & Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422**, 198–207. 40

[177] Stoevesandt, O, Taussig, M. J, & He, M. (2009) Protein microarrays: high-throughput tools for proteomics. *Expert Review of Proteomics* **6**, 145–157. 40

[178] Lahesmaa-Korpinen, A, Jalkanen, S, Chen, P, Valo, E, Núñez-Fontarnau, J, et al. (2011) FlowAnd: Comprehensive Computational Framework for Flow Cytometry Data Analysis. *Journal of Proteomics & Bioinformatics* **4**, 245–249. 40

[179] Schadt, E. E, Turner, S, & Kasarskis, A. (2010) A window into third-generation sequencing. *Human Molecular Genetics* **19**, R227–R240. 40

[180] Clarke, R, Ressom, H. W, Wang, A, Xuan, J, Liu, M. C, Gehan, E. A, & Wang, Y. (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer* **8**, 37–49. 40

[181] Russell, S. J, Norvig, P, Canny, J. F, Malik, J. M, & Edwards, D. D. (2002) *Artificial intelligence: A modern approach*. (Prentice Hall), Second edition. 40

[182] Turing, A. M. (1950) Computing machinery and intelligence. *Mind* **59**, 433–460. 40

[183] Arnaout, R. (2012) Elementary, My Dear Doctor Watson. *Clinical Chemistry* **58**, 986–988. 41